Application of variations of non-linear CCA for feature selection in drug sensitivity prediction

Tolou Shadbahr

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 23.5.2019

Thesis supervisor:

Prof. Juho Rousu

Thesis advisor:

M.Sc. Viivi Uurtio

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Tolou Shadbahr

Title: Application of variations of non-linear CCA for feature selection in drug sensitivity prediction

Date: 23.5.2019 Language: English Number of pages: 6+65

Department of Computer Science

Professorship: Bioinformatics

Supervisor: Prof. Juho Rousu

Advisor: M.Sc. Viivi Uurtio

Cancer arises due to the genetic alteration in patient DNA. Many studies indicate the fact that these alterations vary among patients and can affect the therapeutic effect of cancer treatment dramatically. Therefore, extensive studies focus on understanding these alterations and their effects. Pre-clinical models play an important role in cancer drug discovery and cancer cell lines are one of the main ingredients of these pre-clinical studies which can capture many different aspects of multi-omics properties of cancer cells. However, the assessment of cancer cell line responses to different drugs is faulty and laborious. Therefore, in-*silico* models, which perform accurate prediction of drug sensitivity values, enhance cancer drug discovery.

In the past decade, many computational methods achieved high performances by studying similarity between cancer cell lines and drug compounds and used them to obtain an accurate predictive model for unknown instances. In this thesis, we study the effect of non-linear feature selection through two variations of canonical correlation analysis, KCCA, and HSIC-SCCA, on the prediction of drug sensitivity. To estimate the performance of these features we use pairwise kernel ridge regression to predict the drug sensitivity, measured as $\ln IC_{50}$ values. The data set under study is a subset of Genomics of Drug Sensitivity in Cancer comprise of 124 cell lines and 124 drug compounds.

The high diversity between cell lines and drug compound samples and the high dimension of data matrices reduce the accuracy of the model obtained by pairwise kernel ridge regression. This accuracy reduced by employing HSIC-SCCA method as a dimension reduction step since the HSIC-SCCA method increased the differences among the samples by employing different projection vectors for samples in different folds of cross-validation. Therefore, the obtained variables are rotated to provide more homogeneous samples. This step slightly improved the accuracy of the model.

Keywords: Drug sensitivity, IC_{50}, CCA, KCCA, HSIC-SCCA, pairwise kernel ridge regression, cancer cell lines.

Preface

Foremost, I would like to thank my supervisor, Prof.Juho Rousu, for his supervision at the Department of Computer Science, for his patience, support, and extensive knowledge. My sincere thanks also go to my advisor, Viivi Uurtio for guiding me in the right direction with her comments. I would also like to thank Dr. Sandor Szedmak for his ingenious comments and ideas that help me to learn more. Furthermore, I want to thank the current members and alumni of Prof. Rousu'sresearch group for all their help and friendship. My thanks also go to my friend, Eric Bach, for all constructive discussions that broaden my knowledge and perspectives. Also, I thank all my friends for their support and encouragements. In the end, I want to express my heartfelt thanks to my parents whom I am forever grateful for their unconditional love and support. Lastly, I would like to thank my sister, Taravat Shadbahr, who always lift my spirit with her cheerful smiles.

Espoo, 23.05.2019

Tolou Shadbahr

Contents

Symbols and abbreviations

Symbols

$\mathbf{K}_D$	Kernel matrix for drug compounds
$\mathbf{K}_C$	Kernel matrix for cell lines
$\mathbf{K}_R$	Kernel matrix for Drug sensitivity
λ	Regularization parameter
$\mathbf{I}$	Identity matrix
$\mathbf{1}$	Vector of ones
$\mathbf{C}$	Covariance matrix

Operators

$\otimes$	Kronecker product
$\|\cdot\|_2$	l_2 norm
$\|\cdot\|_1$	l_1 norm
vec	Vectorization operator
$\langle \mathbf{a}, \mathbf{b} \rangle$	Dot product of vectors $\mathbf{a}$ and vector $\mathbf{b}$
$trace$	trace of a matrix

Abbreviations

KRR	Kernel ridge regression
CCA	Canonical correlation analysis
KCCA	Kernel canonical correlation analysis
HSIC-SCCA	Sparse non-linear canonical correlation analysis with HSIC independence criterion
RKHS	Reproducing kernel Hilbert space
PSD	Positive semi-definite
HSIC	Hilbert-Schmidt independence criterion
CV	Cross-validation
RMSE	Root square error

1 Introduction

Many complex diseases, such as cancer, arise due to the alteration in the patient genome. These alterations are responsible for the developments of cancer. Therefore, the responses to treatment can vary widely from a patient to another. This information can be used to tailored treatments to target cancer genetics and improve the survival rate of patients. For example, in the chronic myeloid leukemia, the use of a specific protein as a target of treatment increased the survival rate of the patient by 5-years in 90% of the treated patients. (Garnett et al., 2012; Garnett and McDermott, 2014; Niu and Wang, 2015). These improvements are possible by studies in the genome sequencing and molecular pathology through high-throughput technologies, and bioinformatics and system biology approach Hoelder et al. (2012). However, obtaining a successful anticancer treatment is highly challenging due to the extreme diversity between the genetics of tumors (Lee et al., 2018).

Due to the fact that clinical trials are expensive, time-intensive, and limited to the known drug compounds, the use of pre-clinical models is highly beneficial. These pre-clinical models can be used as biological models to stratify patients and accordingly increase the success rate in the clinical development (Costello et al., 2014; Iorio et al., 2016; Costello et al., 2014). Among all the pre-clinical model, cancer cell lines show an extreme capacity in capturing many different genomics aspects of primary tumors. Additionally, cell lines are renewable sources, and many of genomics data, such as gene expression measurements, for them are publicly available through online databases which are constantly accumulating. Therefore, cell lines are considered well-controlled systems for studying the effect of drug or combination of drugs (Niu and Wang, 2015). The NCI60 cell line panel is the pioneer in the path of using cancer cell lines for studying the link between drug sensitivity and genotype data (Garnett et al., 2012).

However, drug sensitivity data sets for cell lines assays are not complete most of the time due to faulty nature of experiments and measurements machines, and the pre-clinical experiments to obtain these values are time-consuming, laborious and expensive. On the other hand, to achieve a more accurate prediction of drug efficacy on a cell line, increasing the number of samples is highly beneficial. Therefore, an accurate in-*silico* model that can predict the therapeutic response of a drug compound on a cell line, can improve the process of drug discovery (Menden et al., 2013; Cichońska et al., 2018).

Many different machine learning approaches have been suggested for predicting drug sensitivity. These models integrate different sources of information to improve the accuracy of their predictions (Jang et al., 2014). For example, in quantitative structure-activity relationship analysis, chemical and structural features of drug compounds combined with drug sensitivity to obtain a predictive model (Ekins et al., 2007; Ammad-ud din et al., 2016). Furthermore, many high-performance predictive models are built on different genomics measurements combined with the drug sensitivity (Ammad-ud din et al., 2016). Many successful models combined the cell line features with drug compound properties to improve the prediction of drug sensitivity. The works by Ammad-Ud-Din et al. (2014); Costello et al. (2014); Cichonska et al. (2018) are only a few examples. Additionally, the computational models for this regression task vary

from linear methods, such as multivariate linear regression, partial least squares, and principal component regression to non-linear models (Garnett et al., 2012; Barretina et al., 2012). For example, Menden et al. (2013) used neural networks to build a predictive model for obtaining drug sensitivity from features of drug compounds and cell lines. Studies such as Cichonska et al. (2018); Costello et al. (2014) benefits from kernel methods, and in the work of Ammad-Ud-Din et al. (2014) kernel methods combined with matrix factorization to construct the predictive models.

One of the main challenges that predictive models for drug sensitivity values are facing is a large number of variables or features. For instance, each drug compounds can be described by a large number of features, and the gene expression measurements, which are important variables for biological samples, can contain tens of thousands of features. These large numbers of features, make the accurate prediction more difficult. These difficulties are due to the several reasons such as; the prediction function may not converge to the optimal solution in a reasonable time, or more accurate prediction may need a higher number of training data (Liu, 2004).

Among all machine learning methods, kernels are powerful tools in drug sensitivity prediction because of two main points. Firstly, kernels can extract the non-linear patterns between chemical and genomic features, which can be used in the well-known linear learning algorithms. Secondly, kernels overcome the time and memory consumption of the predictive models when the original input is high dimensional. That is, when the number of features is higher than the number of samples such as gene expression in the cell lines, kernels are able to reduce the data input size by embedding the original input data to a matrix of size $n \times n$ where n is the number of samples in the original data set (Cichońska et al., 2018; Costello et al., 2014; Shawe-Taylor et al., 2004).

Additionally, in the high dimensional data sets, such as biological and chemical data sets, dimension reduction for input data is often adapted. In the dimension reduction, it is essential to identify the most relevant variables and the relation between them, and choose the best smaller subset of variables that maximize the prediction accuracy (Cichońska et al., 2018; Ferreira and Purcell, 2008; Tang and Ferreira, 2012; Inouye et al., 2012; Marttinen et al., 2013). These methods enhance the prediction by removing irrelevant variables from the input space and help the model to find the correct predictive function in a shorter time and with smaller training set (Liu, 2004; Ma and Huang, 2008). Among various dimension reduction methods, variations of Canonical Correlation Analysis (e.i CCA) have been adapted successfully for high-dimensional genomics data. In the original CCA, proposed by Hotelling (1936), two sets of variables are studied in order to maximize the correlation between the linear combination of each the variable sets (Hotelling, 1936). Sparse and regularized variations of CCA were introduced to explore the multivariate analysis in the underdetermined setting. Additionally, the exploration of non-linear relations between two data sets has been enhanced by different variations such as Kernel CCA and Deep CCA (Uurtio et al., 2018b; Andrew et al., 2013; Gretton et al., 2008).

The ultimate goal of any in-*silico* model is the accurate prediction of the system under study. These predictions can be classified into two categories. First, the classification prediction e.g. classification of different type of cancers based on their multi-omics properties. Second, the regression prediction such as predicting the binding affinity

between a target protein and a drug compound (Awada et al., 2012; Cichońska et al., 2018). In this study, we are focusing on the prediction of drug sensitivity, that are real values, and fall into the regression prediction category. Generally, in order to combine the strength of the dimension reduction methods with regression prediction, the latent factors extracted by dimension reduction methods can be used for constructing the regression models. This procedure has been successfully adapted in several studies related to the gene expression, such as the paper by Marttinen et al. (2013).

Additionally, kernel ridge regression, a method that combines kernel methods with the ridge regression (i.e. linear least squares with the l2-norm regularization) is used to obtain an optimal regression model in several studies. In order to be able to use the linear least squares cost function for studying data sets with high dimension, the different regularization term, such as LASSO, elastic net, and l_2-norm, are added to the linear least squares cost function in order to prevent the model from over-fitting on the train set and control the trade of between training error and model complexity. Moreover, the systematic studying of the drug sensitivities depends on both cell lines samples and drug compounds. Thus, the better prediction can be achieved by studying the similarities among both instances sets, cell lines samples and drug compounds samples. Therefore, pairwise learning models, that are able to predict based on pairs of instances such as drug compounds and cell lines, are favorable (Cichońska et al., 2018; Menden et al., 2013; Pahikkala et al., 2014).

The objective of this study is to evaluate the effect of various nonlinear Canonical Correlation Analysis (i.e. CCA) techniques as dimension reduction for drug-cell line responses. Therefore, a subset of genomics of drug sensitivity in cancer (GDSC), comprises of 124 drug compounds and 124 cell lines, has been selected to evaluate the accuracy of the obtained model. This data set comprises of 15376 IC_{50} values of cell line responses to different drugs and it is publicly available through Wellcome Sanger institute webpage (Yang et al., 2012). The baseline gene expression values of cell lines, 13321 gene expression has been used as features for cell lines (Ammad-ud din et al., 2016). Additionally, a set of 588 fingerprints has been used as features of drug compounds (Gasteiger and Engel, 2006). The accuracy of obtained features from non-linear canonical correlation, HSIC-SCCA (Uurtio et al., 2018a) and KCCA AKAHO (2001); Bach and Jordan (2002), is evaluated by the accuracy of pairwise kernel ridge regression model obtained by these features.

The chapters of this thesis are as follow; first, we cover biological background relevant to the cell lines responses to the different drug compounds. That is, we explain cell lines, their importance in the drug discovery and their features that can be used for accurate prediction of their responses to different drugs. Additionally, drugs and some of their features that are used for similarity search in chemical databases are explained. Next, we present the computational background and explain the canonical correlation analysis and several variations of it such as regularized and non-linear CCAs. Moreover, the kernel methods and pairwise kernel ridge regression are explained in the computational background. In the material and method section, we focus on our data set. Additionally, the protocols of the experiment are explained in details in this chapter. The result is presented in the result and discussion chapter. In the end, the conclusion of this study has been provided.

2 Biological Background

The somatic mutations in genomes of cells are known to be the emergence of all cancers. These mutations modify the function of proteins that are produced by key cancer genes, thus; not only initiate the cancers but also are responsible for its proliferation. Many of these alterations are known to be important indicators in the determination of the responses to the clinical treatment (Garnett and McDermott, 2014; Iorio et al., 2016; Niu and Wang, 2015). Therefore, the molecularly targeted cancer drug discovery emerged in order to increase the success rate of cancer treatment by revealing potential targets in the cancer pathogenetic drivers and genes. There are many ample shreds of evidence to validate the molecularly targeted therapy (Hoelder et al., 2012).

However, despite recent advances, many more therapeutic options are needed and the process of introducing new drugs is slow due to the high failure rate. There are two main reasons for this fact. Firstly, the genomic heterogeneity is significantly high among tumors, and this includes even those samples that are from the same tissue of origin. Recognizing these differences and targeting a specific cancer gene in the therapy can determine the fate of patients. Secondly, due to a large number of cancer genomes were sequenced, the number of genetic alteration among cancer genomes, both between different tumors and individual cancer, has been increased dramatically. Therefore, finding key driver mutations are incredibly complex and challenging (Hoelder et al., 2012; McDermott and Settleman, 2009). Additionally, validation of new treatments with the clinical trials is time-intensive, expensive and only restricted to the drugs that their safety has been verified. Thus, pre-clinical biological models have been employed in order to provide tractable biological samples that capture molecular features of diseases and their responses to the different treatments as authentic as possible (Costello et al., 2014; Iorio et al., 2016).

Among all pre-clinical biological models, human cancer cell lines are one of the well-established models that are extensively used in the studying of the targeted cancer treatment and cancer biology. Many studies, which focused on the comparisons between cancer cell lines and the original tumors, indicated that cell lines are able to recapitulate majority of diversity and aberrations among 'omic' features of the original tumors (Costello et al., 2014; Goodspeed et al., 2016; Garnett and McDermott, 2014). For examples, comparison between breast cancer tumors and their cell lines indicated the global similarity between gene expression patterns. Additionally, the same resemblance has been observed for the copy number alterations (CNA) between breast cancer tumors and their cell lines (Goodspeed et al., 2016). Furthermore, cancer cell lines are renewable resources with comprehensive multi-omic data publicly available. All these reasons make cancer cell lines invaluable in-*vitro* models in the cancer drug discovery field (Niu and Wang, 2015; Goodspeed et al., 2016).

On the other hand, most approved drugs are considerably small chemical molecules, however; not every chemical compound acquires the desirable physicochemical qualification to be a drug. Additionally, the therapeutic effect of drugs on complex disease such as cancer is affected by the germline and somatic mutation of the patient. That is, a drug with a positive effect on a patient can be completely ineffective on another patient due to the different genome background of the patient (Cichońska et al., 2018;

Pan et al., 2013; Niu and Wang, 2015). This fact increases the need for searching between chemical compounds based on their similarity in order to find drug-like compounds with a higher rate of success in the pre-clinical and clinical trails (Ammad-ud din et al., 2016; Ammad-Ud-Din et al., 2014).

In the following, first a compact review of cell lines history is given. Then the drug sensitivity of cell line is explained as a metric for studying the effectiveness of a drug on a cell line. Then, the gene expression concept, as one of the main feature extracted and frequently studied in cancer drug discovery, is explained. At the end of this section, drugs and their features, which can be used for similarity search and studying the therapeutic effect of drugs and finding new drug compounds, is given.

2.1 Cell Lines as Biological Samples

Cell and tissue cultures have made a large impact on biological science since their starting point in 1885. Since then there have been several milestones in the cell and tissue cultures history. One of the most important milestones is using human tissue for the first time in 1898. Soon after, attempts for longer preservation of a culture experiment have been made. The continuation of the functionality of cells in-*vitro* is illustrated by these experiments and were important in the development of general techniques in tissue cultures (Langdon, 2010).

In 1951, the first human cancer continuous cell line has been developed, and in the two decades, the 1950s and 1960s many studies by different scientists have been made regarding the nutrition of cells in the media. Since the 1970s, an extensive number of models for a different type of cancers have been provided by a huge number of cell lines (Langdon, 2010). In 1980, the National Cancer Institute (NCI) established the NCI-60 project, which comprises of 60 human cancer cell lines extracted from 9 different types of cancers. In this project, extensive omics data for cell lines have been collected and stored in order to integrate the multi-omics data into the targeted cancer drug discovery (Niu and Wang, 2015; Goodspeed et al., 2016).

Now by considering this history, we will explain what is a cell culture and a cell line. The term cell culture indicates the act of removing cells from animals or plants to observe their growth in a sufficient artificial environment, (Langdon, 2010; Scientific, 2015). The culture is called primary culture when we initializing it by a sample from an individual. After transferring the diluted prime culture to a new container, it refers to a cell line. This transferring procedure is called subculture or passage. If the cell line is able to expand its population indefinitely, it is called a continuous line. On the hand, when the cell line is capable of a limited number of doubling in the population is called the finite line, (Langdon, 2010).

In order to evaluate the therapeutic responses of cancer cell lines, the drug dose-response of cultured cells (i.e. cell lines) need to be measured. Therefore; drug response assays are used to evaluate the effectiveness and performance of tested compounds. That is, the drug effectiveness over a range of concentrations has been evaluated and reported by drug response assays. There are several metrics used to report drug efficacy. Among all these metrics IC_{50} is one of most commonly used drug response metric (Sebaugh, 2011; Niepel et al., 2017; Hafner et al., 2016). In order to obtain the IC_{50} for

anti-cancer drugs, cell lines are exposed to a range of concentrations of a drug over a specific period of time (e.g, 72 hours). For cell lines the reduction in the cell counts indicates the cell growth inhibition, therefore; the number of viable cells is counted after the indicated period of time. Specifically, cell counts in the samples treated by different concentrations of a drug divided by the number of the cells in the untreated samples (i.e. control sample) are fitted to a sigmoidal curve. The IC_{50} is indicated by the concentration of the drug which the number of cells in the treated sample is half of the control sample (Hafner et al., 2016; Van Meerloo et al., 2011).

In the drug response studies, multi-omics data, such as gene expression, are combined with drug response measurements. Now that a brief overview of the cell culture and cell lines is given, in the next section we will explain the meaning of gene expression as an important source of information in biological study. We will start with a review of what is a gene, what is gene expression, and how it can be measured.

2.1.1 Gene Expression

Our body consists of an enormous number of cells, and each cell contains a nucleus. The complete human genome can be found in almost every nucleus of human cells. In the human body, the genome defines all cellular structures and activities. The genome consists of 23 chromosomes pairs. Two huge strands of deoxyribonucleic acid (DNA) intertwine together and formed a chromosome. Each DNA strand consists of nucleotides' sequences and each nucleotide build of a phosphate group, a deoxyribose sugar, and one of the four nitrogen bases (e.g. Adenine (A), Thymine (T), Cytosine (C), and Guanine (G)) (Karp, 2009; Schwender et al., 2006).

Segments of a DNA are called genes. Genes are the source of information for constructing proteins. All structures and activities of a cell are defined by proteins; therefore, it is important to study and observe the translation of genes to proteins. Steps of this procedure are as follow; at first, a single strand of DNA transcribes to an mRNA and leaves the nucleus. Secondly, each codon (3-mer) of this mRNA is translated to one of 20 possible amino acids and they shape a chain of amino acids. At last, this chain folds into a protein. This procedure is called gene expression and it is known as *Central Dogma of Molecular Biology* (Karp, 2009; Schwender et al., 2006). Figure 1 illustrates the steps of gene expression procedure.

The expression of genes can be alerted by several factors such as alternative splicing, binding different transcription factor proteins to the regulatory or promoter regions of a DNA, and DNA differences between humans. This means a single gene can be coded for several proteins. Genetic variation, such as deletions or substitutions of bases, are called mutation, variant or variation. Several diseases are caused by an inherited mutation in a single gene. Lethal diseases, such as cancer, are known to have many factors, such as genetic variants and environmental factors. Therefore, it is of high importance to study several genes and their expression (Schwender et al., 2006; Schwender, 2007).

The distribution and state of proteins in each cell defines the properties and functions of the cells. Therefore, By measuring the protein's abundance in a cell, the functionality of this cell can be studied. However, monitoring the expression level of a protein is more complex compared to mRNA. Additionally, based on the Central Dogma of

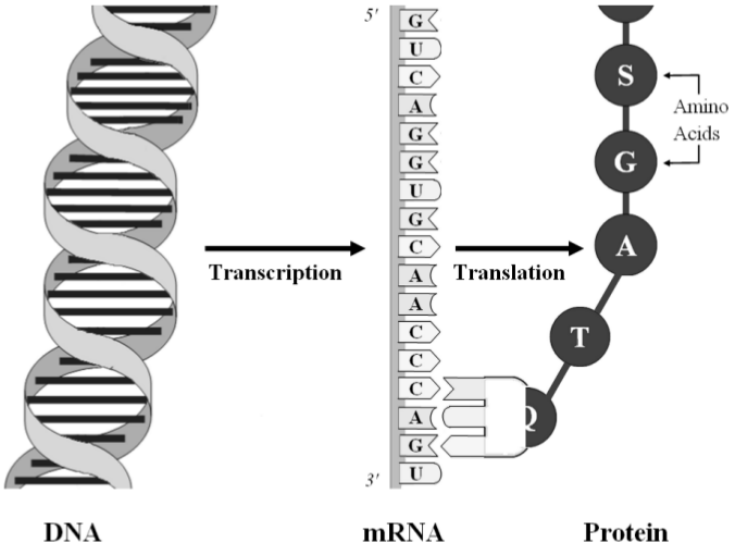


Figure 1: Central Dogma of the Molecular Biology (Schwender, 2007)

Molecular Biology (Karp, 2009), mRNA measurements also can reveal the cell function. DNA microarrays are used for measuring the level of expressions for thousands of genes in a cell (Schwender et al., 2006; Schwender, 2007).

There are a variety of microarrays, but most of them have the same basic principle that is called complementary base-pairing. The amount of specific mRNA, and consequently the level of expression for the related gene, is measured as follow;

1 a batch of synthesis complementary DNA probe (i.e. cDNA) corresponding to the mRNA is placed on the microarray.

2 the desired mRNAs are labeled with a fluorescent dye and fragmented into the pieces of 25-500 bases.

3 the labeled mRNAs are passed over the microarray to hybridized with the probs on the microarray.

4 the non-bounded mRNAs are washed away and the array is scanned to measure the expressed mRNAs. The expression of a specific mRNA is proportional to the intensity of the fluorescence.

In the aforementioned procedure, each gene (e.g. mRNA) is presented by two types of cDNA (oligonucleotide) probes. Perfect matches (PM) oligo (short form of oligonucleotide), that are used to measure the abundance of the specified mRNA, and Mismatches (MM), that are helpful for correcting the background noise and non-specific binding, for more details regarding the obtaining the gene expressions and pre-processing procedure (Schwender, 2007). In some type of microarrays methods, specially cDNA microarray, two types of mRNA samples are used. One sample is identical for all arrays and it is called the control sample. The other sample refers to the

mRNA of our interest (e.g. extracted from a cancer tumor sample). Each of these mRNA sequences is labeled by the different dye, green and red. The number of relative gene expressions (over or underexpression) can be calculated by log_2 ratio between the intensity of these two fluorescence dyes (Schwender et al., 2006; Schwender, 2007).

In the next part, we go through to a brief history of advances in the drug discovery. Then, we talk about how features of chemical compounds can be presented, stored and used by different computer programs.

2.2 Drug Compounds

After introducing high throughput screening (HTS) in 1980, that enhanced the procedure of testing different drug compounds on a drug target, the demand for new drug compounds has increased by biologists. Therefore, a new procedure to construct an extensive array of compounds by systematically using building blocks of drug compounds (different types of reagents), combinatorial chemistry (CC), has been adopted by chemists (Xu and Hagler, 2002).

By adopting these new technologies, there were high hopes to facilitate the discovery of new drugs. However, it was not the case due to the fact not all combinations of different building blocks of different drug compounds can result in a drug candidate. In the next attempt to provide more drug candidates, the chemically diverse compound libraries were introduced. These libraries tried to distinguish between the chemical components by their structures. Different computational approaches, such as classification algorithms and structural similarity, have been innovated and applied in the construction of these libraries. These computational methods are ingredients of cheminformatics discipline (Xu and Hagler, 2002). By adopting these libraries, it has been indicated that to have a drug candidate, the chemical compound should meet several conditions such as molecular weight, hydrophobicity, number of rotatable bonds, and must not contain undesired substructure such as toxicophores (Rhodes et al., 2000).

Thus, in order to reinforce searching for the aforementioned conditions, using computers to search the chemical data and databases for structures and substructures has evolved. Structures of chemical components can affect the activities and properties of those components. Due to this fact, distinctive approaches were introduced to store and retrieve structures of a chemical compound. There are several approaches to represent a chemical component. Even a single component can be presented in many different ways. Therefore, many attempts were made to find a succinct approach to represent the chemical compound. Graph pretension and linear notations are good examples of the different structural representation (Xu and Hagler, 2002; Leach and Gillet, 2007; Gasteiger and Engel, 2006).

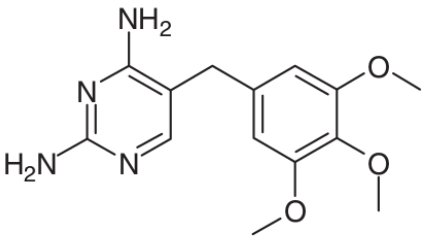
2.2.1 SMILES and Fingerprints in Chemical Compound

There are several linear notations methods that represent the chemical structures by using strings of letters and a set of rules. Simplified Molecular Input Line Entry Specification (SMILES), introduced by David Weininger, is one of the most popular methods for linear notation of chemical structures (Weininger, 1988; Xu and Hagler,

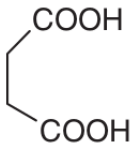
2002; Leach and Gillet, 2007). The comprehensive use of SMILES is because most SMILES strings can be written by simple rules. Some of these rules are given below as an example;

- Atomic symbols have been used for presenting the atoms.
- Aliphatic atoms are presented by Upper case letters, and aromatics are indicated by lower case symbols.
- Hydrogen atoms are not presented normally.
- Double bonds are indicated by “=”, and triple bonds by “#”
- Single and aromatic bonds are not illustrated except in some special cases where they are presented by "-”.
- The rings are indicated by giving an integer number to the two atoms where the ring is broken.
- The starting point of a branch is illustrated by left-hand brackets, and a right-hand bracket indicates that all atoms of that branch are seen.

Additionally, in order to construct the SMILES for a chemical molecule, we need to "walk” through the chemical compound structure in such a way that all the atoms are visited once (Leach and Gillet, 2007; Weininger, 1988). In Figure 2 two examples of SMILES are presented for two different chemical compounds.



trimethoprim: COc1cc(Cc2cnc(N)nc2N)cc(OC)c1OC



succinicacid:
OC(=O)CCC(=O)O

Figure 2: SMILES presentation of two chemical compound; Trimethoprim and Succinicacid (Leach and Gillet, 2007).

SMILES and other linear notation methods have been used to assist the structural search in computer databases and retrieve information. Additional to the structural search, substructure searches are applied widely to identify a certain substructure in all molecules of a database. However, the substructure search is often a slow procedure. The molecule screens are used to accelerate the search and avoid the atom-by-atom matching. In the molecular screening method, a bitstring has been provided for both molecule and query substructure. The bitstring is a sequence of "0”s and "1”s that indicates the presence or absence of a specified feature structure (a pre-defined substructure) (Leach and Gillet, 2007,?; Xu and Hagler, 2002).

Pre-defined substructures may not be efficient and contain some biases when the structures are not seen before. Thus, it is necessary to use systematical approaches to generate substructures. The negative side effect of using systematical substructures generation is that the cost of storage and computation increases. To overcome this side effect, the bit-maps were used. In the bit-maps, a list of substructures is represented by a bit, and same as bitstrings the "0"s and "1"s are used to illustrate the presence or absence of a structure. These bit-maps are known as fingerprints (Leach and Gillet, 2007; Gasteiger and Engel, 2006; Xu and Hagler, 2002). In Figure 3 two examples of fingerprint presentations of two different chemical compounds are given.

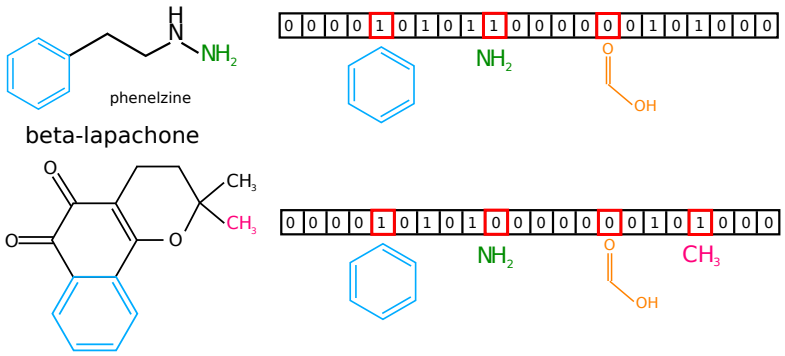


Figure 3: Finger print presentation of two chemical compound; Pheneizine and Beta-lapachone

There are many different fingerprints, and this variety is due to the different substructures that have been used. Daylight fingerprints and MDL fingerprints are two examples out of many existing fingerprints, there are many more fingerprints listed in the papers Gasteiger and Engel (2006); Leach and Gillet (2007); Xu and Hagler (2002); Eckert and Bajorath (2007). The similarity between two different chemical structures can be computed by fingerprints in a shorter time and by less computational efforts. Additionally, the same number of fingerprints can be provided for two structures that are different in the number of atoms and bonds (Xu and Hagler, 2002; Gasteiger and Engel, 2006; Leach and Gillet, 2007)

The similarity measured based on the fingerprints has been used for identifying the structures that have similar properties. This simple fact can be used to study the biological activities of unknown chemical molecules that have similar structures to a molecule with known biological activities. That is, an unknown chemical molecule similar to a molecule with known properties is likely to have the same properties (Willett, 2006). This fact inspires the rise of machine learning and data mining techniques in studying similarities between chemical compounds. In other words, obtaining a model that is capable of mapping several molecular descriptors to different important biological attributes, and adapting this model predict properties of unknown compounds. There has been a wide range of machine learning methods employed to study and analysis chemical molecules, such as support vector machines (SVM), decision tree, artificial

neural networks, and many others (Lavecchia, 2015).

In this chapter, we presented concepts of cell lines, drug compounds and their possible feature representations, e.g. gene expressions and fingerprints. Moreover, we explained the importance of cell lines as a patient avatar in cancer targeted drug discovery. That is, a cell line can be used as an in-*vitro* model with the similar features as the patient to investigate the effect of different drugs or drug-like compounds on the patient or a type of cancer (Niu and Wang, 2015; Garnett and McDermott, 2014). A therapeutic effect of a drug on the cell lines is indicated by the drug sensitivity and the accurate prediction of this value is important. However, testing different drug compounds on different cell lines is time taking and expensive. Therefore, an in-*silico* model that can accurately predict drug sensitivity is highly beneficial Menden et al. (2013).In the next chapter, we study several machine learning methods that can be used for similarity search in the drug compounds and cell lines data sets and constructing predictive models for drug sensitivity.

3 Computational Background

In this chapter, we explain the computational methods that have been used in this thesis. First, we start by introducing kernel methods. The kernel ridge regression and pairwise kernel ridge regression are explained, secondly. In the next step, the basics of CCA and its regularisation variation (i.e. RCCA) are given. This basic introduction is followed by describing the non-linearity relation analysis and sparsity in the CCA by explaining three different methods; sparse CCA (i.e. SCCA), Kernel CCA (i.e. KCCA), and HSIC-SCCA. In the end, the cross-validation technique for model selection and model evaluation is explained.

The following notation has been used through this thesis. The matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ indicates the sets of measurements over n samples (i.e instances or data points). The number of features (i.e. variables) is denoted by p. The samples are from an input space denoted by $\mathbb{R}^{n \times p} \in \mathscr{X}$. In this matrix, the samples are stored as rows vectors in the matrix and the columns of this matrix correspond to the features. The sample i is presented as column vector $\mathbf{x}^i$, where $\mathbf{x}^i \in \mathbb{R}^p \in \mathscr{X}$. Matrix $\mathbf{Y} \in \mathbb{R}^n$ contains the output (i.e. response) measurements of the instances and its output space is indicated by $\mathscr{Y}$.

3.1 Kernel Methods

The kernel methods have been employed for the first time in the field of pattern recognition by Aizerman (1964). However, after many years, it re-appeared in the concept of machine learning to extend the Support Vector Machine method to a non-linear classifier (Melzer et al., 2001; Bishop, 2006). Kernel methods are used when a function consists of a non-linear combination of features only can obtain an accurate estimation of a label value, either in classification or regression task. Before defining the Kernel methods, we explain several basic definitions first.

Definition 3.1. Inner product: The inner product (i.e dot or scalar product) of two input samples $\mathbf{x}, \mathbf{z} \in \mathbb{R}^P$ can be calculated by $\langle \mathbf{x}, \mathbf{z} \rangle = \sum_{i=1}^{p} x^i z^i$, where $\mathbf{x}$ and $\mathbf{z}$ refer to individual samples in the input space (Shawe-Taylor et al., 2004).The obtained vectors from inner product operation construct a gram/kernel matrix and they belong to the inner product space when they contain only real values (Melzer et al., 2001; Bishop, 2006; Shawe-Taylor et al., 2004).

Definition 3.2. Gram matrix: A gram matrix of a given set of vectors, $S = \{\mathbf{x}^1, \ldots \mathbf{x}^l\}$ is a $l \times l$ matrix $\mathbf{G}$ with entries equal to $\mathbf{G}^{ij} = \langle \mathbf{x}^i, \mathbf{x}^j \rangle$ (Shawe-Taylor et al., 2004).

Now, by understanding these basic definitions, we can explain kernel methods.

Definition 3.3. Kernel methods: Kernel Methods induce the non-linearity to a linear function by mapping the features (e.g. variables) of input data set into a nonlinear feature space. Assume samples are indicated by $\mathbf{x}^i \in \mathbb{R}^P$ where $i = 1, \ldots, n$ is the number of samples. The Equation 1 represent this mapping (Shawe-Taylor et al., 2004).

$$\phi : \mathbf{x} = (x_1, x_2, \cdots, x_p) \in \mathbb{R}^P \rightarrow \phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \cdots, \phi_s(\mathbf{x})) \in \mathbb{R}^s, \quad s > p. \tag{1}$$

A linear function can be presented as $\mathbf{w}^T\mathbf{x}^i$ where $\mathbf{w} \in \mathbb{R}^p$ is a coefficient vector. The linear combinations of these obtained mapped features, $\mathbf{w}^T\phi(\mathbf{x})$, can be used as a model for the prediction task. This new formulation of the predictive model can be expressed as an inner product of the given feature maps. In the Equation 2, a kernel function is presented by the inner product of the feature space mapping $\phi(x)$.

$$k(\mathbf{x}^i, \mathbf{x}^j) = \langle \phi(\mathbf{x}^i), \phi(\mathbf{x}^j) \rangle = \phi(\mathbf{x}^i)^T \phi(\mathbf{x}^j), \tag{2}$$

where $k(\mathbf{x}^i, \mathbf{x}^j) \in \mathbb{R}^{n \times n}$. A kernel measures the similarity between a pair of samples $\mathbf{x}^i$ and $\mathbf{x}^j$. However, not every similarity measure can be considered as a valid kernel. In order to a similarity measure yields, a valid kernel the positive semi-definite (PSD) criteria should be met.

Definition 3.4. Positive semi-definite: A symmetric matrix $\mathbf{X}$ is PSD if $\mathbf{u}^T\mathbf{X}\mathbf{u} \geq 0$ for all non-zero vector $\mathbf{u} \in \mathbb{R}^n$. In the PSD matrices, all the eigenvalues are non-negative. Additionally, a matrix $\mathbf{X}$, is symmetric if $\mathbf{X} = \mathbf{X}^T$ (Shawe-Taylor et al., 2004; Cichońska et al., 2018).

Two examples of valid kernels are presented in the Equation 3.

$$\begin{aligned} \text{linear kernel:} \quad & k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}' \\ \text{Gaussian kernel:} \quad & k(\mathbf{x}, \mathbf{x}') = \exp(\frac{-\mathbf{x}^T\mathbf{x}}{2\sigma^2}) \exp(\frac{\mathbf{x}^T\mathbf{x}'}{\sigma^2}) \exp(\frac{-(\mathbf{x}')^T\mathbf{x}'}{2\sigma^2}) \end{aligned} \tag{3}$$

In order to better understanding of the Equation 2 a simple example is given in the following. For the sake of simplicity, the input dimension is considered to be two in this example.

Assume a given data point $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$. The following feature map can transform the given data point in a two-dimensional space into a three-dimensional space.

$$\phi : \mathbf{x} = (x_1, x_2) \mapsto \phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2) \in \mathbb{R}^3.$$

Thus, a linear function in the corresponding feature space can be expressed as following:

$$f(\mathbf{x}) = w_{11}x_1^2 + w_{22}x_2^2 + w_{12}\sqrt{2}x_1x_2,$$

where $w_{11}, w_{22}, w_{21} \in \mathbb{R}$ are the model parameters. As can be seen, this linear function in the feature space is equivalent to a quadratic function in the input space. Additionally, the inner product of feature maps for two given data points in the feature space can be illustrated as follow;

$$\begin{aligned} \langle \phi(\mathbf{x}^i), \phi(\mathbf{x}^j) \rangle &= \langle (x_1^{i^2}, x_2^{i^2}, \sqrt{2}x_1^i x_2^i), (x_1^{j^2}, x_2^{j^2}, \sqrt{2}x_1^j x_2^j) \rangle \\ &= x_1^{i^2} x_1^{j^2} + x_2^{i^2} x_2^{j^2} + 2x_1^i x_2^i x_1^j x_2^j \\ &= (x_1^i x_1^j + x_2^i x_2^j)^2 = \langle \mathbf{x}^i, \mathbf{x}^j \rangle^2, \end{aligned}$$

where i,j correspond to the different samples (i.e data points) in the input space. Thus, $k(\mathbf{x}^i, \mathbf{x}^j) = \langle \mathbf{x}^i, \mathbf{x}^j \rangle^2$ is a kernel function (Shawe-Taylor et al., 2004).

The kernel methods not only enhance the construction of non-linear predictive models but also reduce the complexity of the prediction model when the number of instances is smaller than the number of variables. That is due to the fact that the prediction no longer depending on the feature size but depending on the sample size (Melzer et al., 2001; Bishop, 2006; Shawe-Taylor et al., 2004).

Definition 3.5. Hilbert space: An inner product space, such as $\mathscr{F}$, that is separable and complete is a Hilbert space. Assume $h_1, h_2, \ldots, h_N$ are elements of Hilbert space $\mathscr{F}$, and a $\epsilon > 0$. If a finite set of elements exists in the space $\mathscr{F}$, for all $h \in \mathscr{F}$, the condition $\min_i \|h_i - h\| < \epsilon$, then the space $\mathscr{F}$ is separable. Additionally, if all Cauchy sequence of elements of $\mathscr{F}$, $\{h_n\}_{n\geq 1}$, converge to an element $h \in \mathscr{F}$. Then we say the space $\mathscr{F}$ is complete (Shawe-Taylor et al., 2004).

Definition 3.6. Centering of a gram matrix: In order to move the origin of feature space to the center of instances in the given data set the centering technique can be employed. A centered gram matrix is indicated by $\hat{K} \in \mathbb{R}^{n\times n}$ in the Hilbert space $\mathscr{F}$, and obtained from the Equation 4.

$$\hat{\mathbf{K}} = \mathbf{HKH}, \tag{4}$$

where $\mathbf{H}_{ij} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ and $\mathbf{1}$ is a vector of ones with size n. Additionally, the sum of the norms of samples is minimal in the center (i.e origin). The centering of a gram matrix can affect the sum of eigenvalues of the corresponding gram matrix since The sum of eigenvalues of a gram matrix is minimized by centering the Gram matrix (Shawe-Taylor et al., 2004).

Definition 3.7. Reproducing kernel Hilbert space (RKHS): Assume a Hilbert space $\mathscr{H}_f$ and a point x in this space. Additionally, imagine a point evaluation function $\delta : \mathscr{H}_f \to \mathbb{R}$, which maps any function $\boldsymbol{f} \in \mathscr{H}_f$ to $f(x) \in \mathbb{R}$. The Hilbert space $\mathscr{H}_f$ is a reproducing kernel Hilbert space if the point evaluation function is continuous. All reproducing kernel Hilbert spaces defined by a positive definite kernel Gretton et al. (2005b,a). One of the main properties of RKHS is that the norm of a given RKHS space $\mathscr{H}_k$ can be calculated by representer theorem. We will define this theorem in the following.

Definition 3.8. Representer theorem(Schölkopf et al., 2001): Assume a non-empty set $\mathbf{X}$ from input space $\mathscr{X}$, a positive definite real-value kernel k on the space $\mathscr{X} \times \mathscr{X}$. Moreover, a training data set comprises of input and responses pairs, $(x^1, y^1), \ldots, (x^m, y^m) \in \mathscr{X} \times \mathscr{R}$, a strictly monotonically increasing real value function $\lambda \in [0, \infty)$, a class of functions such as $\mathscr{F} = \{f \in \mathscr{R}^{\mathscr{X}} \,|\, f(\cdot) = \sum_{i=1}^{\infty} \beta_i k(\cdot, z_i), \beta_i \in \mathscr{R}, z_i \in \mathscr{X}, \|f\| < \infty\}$, and an arbitrary loss function $\mathscr{L} : (\mathscr{X} \times \mathscr{R}^2)^m \to \mathbf{R} \cup \infty$ are given. The operation $\|\cdot\|$ refers to the norm of a RKHS space $\mathscr{H}_k$ corresponding to k and can be calculated by the Equation 5.

$$\|\sum_{i=1}^{\infty} \beta^i \mathbf{k}(\cdot, z^i)\|^2 = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \beta^i \beta^j k(z^i, z^j), \tag{5}$$

where, $z_i \in \mathscr{X}$ is arbitrary unseen samples, $\beta_i \in \mathbf{R}$ are coefficients, and $i \in \mathbf{N}$ is index of unseen samples. Then any function $f \in \mathscr{F}$ minimizing the regularized loss function,

$\mathscr{L}((x_1, y_1, f(x_1)), \cdots, (x_m, y_m, f(x_m))) + \lambda(\|f\|)$, on the train set can be presented by the Equation 6.

$$f(\cdot) = \sum_{i=1}^{m} \alpha^i k(\cdot, x^i). \tag{6}$$

The $\alpha \in \mathbb{R}^m$ is a vector of coefficients called a dual variable. The variable m indicates the number of instances in the training data set (Schölkopf et al., 2001).

3.2 Kernel Ridge Regression and Pairwise KRR

In the prediction task, the aim is to obtain an optimal model based on the labeled training data which yields the minimum generalization error, that is, error on the test data set. The labeled training data refers to the samples that their labels (i.e. responses) are known. Assume the data matrix $\mathbf{X}$ as input data from space $\mathscr{X}$ and matrix $\mathbf{Y}$ denotes the responses from the space $\mathscr{Y}$. The training data set defines as $D = \{(\mathbf{x}^i, y^i)\}_{i=1}^n$ where sample pair $\mathbf{x}^i \in \mathscr{X}$ and $y^i \in \mathscr{Y}$ are sampled from the underlying unknown joint distribution. A set of models, $\mathscr{F}$, maps the input space to the output space, $\mathscr{F} = \{f : \mathscr{X} \mapsto \mathscr{Y}\}$. In order to obtain the optimal model, f, a loss function that measures the empirical error, the error between the prediction responses and true responses in the training data set, is employed. This loss function is defined as $\mathscr{L} : \mathscr{Y} \times \mathscr{Y} \mapsto \mathbb{R}$ and the empirical risk of the loss function can be calculated by the Equation 7.

$$R(f) = \sum_{i=1}^{n} \mathscr{L}(y^i, f(\mathbf{x}^i)). \tag{7}$$

Therefore, the optimal model f minimises the empirical error R, that is, $\arg\min_{f \in \mathscr{F}} R(f)$ (Cichońska et al., 2018).

However, minimizing the empirical error R given in the Equation 7 on a specified training data set can result in overfitting. That means the optimal function f with very low empirical error can yield high generalization error on unseen data. This situation can happen due to the noise in the training data set. Therefore, a regularization term is added to the Equation 7. The obtained optimization problem is given in Equation 8.

$$\arg\min \sum_{i=1}^{n} \mathscr{L}(y^i, f(\mathbf{x}^i)) + \lambda\omega(f), \tag{8}$$

where $\omega(f)$ indicates a function and its values raises as the complexity of the model increases. The $\lambda \geq 0$ refers to the regularization parameter and it controls the tradeoff between the model complexity and the empirical error. The small norm function is used often as a regulariser function (Cichońska et al., 2018).

The ridge regression, that combines the squared loss function with $l2$-norm regularise, has been used regularly in the regression prediction. Assume a linear model of the form $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle = \mathbf{w}^T\mathbf{x} = \sum_{l=1}^{t_x} w_l x_l$, where $\mathbf{w} \in \mathbb{R}^{t_x}$ is a vector of model coefficients obtained by minimising the empirical risk. Additionally, t_x is the number of the samples in the training set. The squared loss function is $\mathscr{L}(y^i, f(\mathbf{x}^i)) = (y^l - f(\mathbf{x}^l))^2$. Therefore,

the ridge regression is defined as Equation 9.

$$\arg\min_{\mathbf{w}} \sum_{l=1}^{t_x} (y^l - \langle \mathbf{w}, \mathbf{x}^\mathbf{l} \rangle)^2 + \lambda \|\mathbf{w}\|^\mathbf{2} = \arg\min_{\mathbf{w}} \langle \mathbf{y} - \mathbf{X}\mathbf{w}, \mathbf{y} - \mathbf{X}\mathbf{w} \rangle + \lambda \langle \mathbf{w}, \mathbf{w} \rangle. \tag{9}$$

The solution for the optimal coefficient of the model can be calculated by derivatives of the Equation 9 with respect to $\mathbf{w}$ and setting them equal to zero vector. The optimal $\mathbf{w}$ is calculated by the Equation 10.

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{tx})^{-1}\mathbf{X}^T\mathbf{y}, \tag{10}$$

where $\mathbf{I}_{tx}$ refers to the identity matrix of the size $\mathbb{R}^{t_x \times t_x}$ (Cichońska et al., 2018; Saunders et al., 1998).

3.2.1 Kernel Ridge Regression

Kernels can be used in ridge regression to induce non-linearity to the model. The optimal prediction function for the Kernel ridge regression can be obtained by representer theorem, presented in (Schölkopf et al., 2001; Kimeldorf and Wahba, 1971).

The reprensenter theorem can be used in order to obtain the function f that minimized regularized empirical error. Thus, the function f in the Equation 8 can be presented in the Equation 11.

$$f(\mathbf{z}) = \langle \mathbf{w}, \mathbf{z} \rangle = \sum_{i=1}^{m} \boldsymbol{\alpha}^i \langle \mathbf{x}^i, \mathbf{z} \rangle = \sum_{i=1}^{m} \boldsymbol{\alpha}^i k_x(\mathbf{x}^i, \mathbf{z}) = \boldsymbol{\alpha}^T \mathbf{k}_x, \tag{11}$$

where $\mathbf{k}_x$ is a vector comprises of kernel values between samples of the training set, $\mathbf{x}^i$, and test samples, z. Therefore, coefficient vector $\mathbf{w}$ can be obtained as a linear combination of training samples. This is presented in the Equation 12.

$$\mathbf{w} = \sum_{i=1}^{m} \boldsymbol{\alpha}^i \mathbf{x}^i = \mathbf{X}^T \boldsymbol{\alpha}. \tag{12}$$

Additionally, the Equation 9 can be rewrite as minimizing problem of the form $\sum_{i=1}^{m} \zeta^{i2} + \lambda \|\mathbf{w}\|^2$ with constraints $\zeta^i = y^i - \mathbf{w}^T\mathbf{x}^i$ and $i = 1, \ldots, m$. The Lagrange multipliers α^i can be used in order to obtain $\sum_{i=1}^{m} \zeta^{i2} + \lambda \|\mathbf{w}\|^2 + \sum_{i=1}^{m} \alpha^i (y_i - \mathbf{w}^T\mathbf{x}^i - \zeta^i)$. The solution for coefficient vector $\mathbf{w}$ can be obtained by differentiating the mentioned equation with respect to $\mathbf{w}$ and it is equal to $\mathbf{w} = \frac{1}{\lambda} \sum_{i=1}^{m} \alpha^i \mathbf{x}^i$ (Saunders et al., 1998).

The solution to kernel ridge regression (i.e. ridge regression in the dual space) can be acquired by reforming the Equation 10 into $\mathbf{w} = \lambda^{(-1)}\mathbf{X}^T(y - \mathbf{X}\mathbf{w}) = \mathbf{X}^T\boldsymbol{\alpha}$. The closed form solution presented in the Equation 13.

$$\begin{aligned} \boldsymbol{\alpha} &= \lambda^{(-1)}(y - \mathbf{X}\mathbf{w}), \\ \lambda\boldsymbol{\alpha} &= (y - \mathbf{X}\mathbf{X}^T\boldsymbol{\alpha}), \\ (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{T}_n)\boldsymbol{\alpha} &= y, \\ (\mathbf{K}_x + \lambda\mathbf{I}_n)\boldsymbol{\alpha} &= y, \\ \boldsymbol{\alpha} &= (\mathbf{K}_x + \lambda\mathbf{I}_n)^{-1} y, \end{aligned} \tag{13}$$

where $\mathbf{K}_x = \mathbf{x}\mathbf{x}^T$ refers to the kernel matrix comprises of the inner products of the training samples. The prediction for the test instance $\mathbf{z}$ can be calculated by the Equation 11 when the $\boldsymbol{\alpha}$ is computed on the training set (Cichońska et al., 2018; Saunders et al., 1998).

3.2.2 Pairwise Kernel Ridge Regression

In the case of pairwise learning, each instance comprises of two individual parts, which are presented by separate features. Therefore, the training samples can be presented as $D = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{y}) \in (\mathscr{X}_1 \times \mathscr{X}_2 \times \mathscr{Y})$, where $\mathbf{X}_1 \in \mathbb{R}^{n \times p}$ and $\mathbf{X}_2 \in \mathbb{R}^{m \times q}$ are feature matrices of part one and part two respectively. The vector $\mathbf{y} \in \mathbb{R}^N$ stores the labels for sample pairs and $N \leq p \cdot q$. Thus, the goal in the pairwise learning is to obtain an optimal prediction function $f : \mathscr{X}_1 \times \mathscr{X}_2 \mapsto \mathscr{Y}$ that captures the relationship between pairs of samples, $(\mathbf{X}_1, \mathbf{X}_2)$, and their labels y (Cichońska et al., 2018; Pahikkala et al., 2013).

By adopting a pairwise kernel k in the KRR, the KRR can be used for a pairwise learning problem. The Kronecker product pairwise kernel is the most favorable choice for measuring the similarity between sample pairs. The Kronecker product of two kernels, $k_1 : \mathscr{X}_1 \times \mathscr{X}_1 \mapsto \mathbb{R}$ and $k_2 : \mathscr{X}_2 \times \mathscr{X}_2 \mapsto \mathbb{R}$ is given in the Equation 14.

$$k\left((\mathbf{x}_1^i, \mathbf{x}_2^i), (\mathbf{x}_1^j, \mathbf{x}_2^j)\right) = k_1(\mathbf{x}_1^i, \mathbf{x}_1^j).k_2(\mathbf{x}_2^i, \mathbf{x}_2^j), \tag{14}$$

where, $\mathbf{x}_1^i, \mathbf{x}_1^j$ refer to different samples in data matrix $\mathbf{X}_1$, and $\mathbf{x}_2^i, \mathbf{x}_2^j$ indicate individual samples in data matrix $\mathbf{X}_2$. That is, the pairwise kernel $\mathbf{K} \in \mathbb{R}^{n \times n}$ between two kernels $\mathbf{K}_1 \in \mathbb{R}^{n \times n}$ and $\mathbf{K}_2 \in \mathbb{R}^{m \times m}$ is a block matrix and contains all possible products of samples in both kernel. The Equation 15 presents the Kronecker product in more details.

$$\mathbf{K} = \mathbf{K}_1 \otimes \mathbf{K}_2 = \begin{pmatrix} k_1(\mathbf{x}_1^1, \mathbf{x}_1^1)\mathbf{K}_2 & k_1(\mathbf{x}_1^1, \mathbf{x}_1^2)\mathbf{K}_2 & \cdots & k_1(\mathbf{x}_1^1, \mathbf{x}_1^n)\mathbf{K}_2 \\ k_1(\mathbf{x}_1^2, \mathbf{x}_1^1)\mathbf{K}_2 & k_1(\mathbf{x}_1^2, \mathbf{x}_1^2)\mathbf{K}_2 & \cdots & k_1(\mathbf{x}_1^2, \mathbf{x}_1^n)\mathbf{K}_2 \\ \vdots & \vdots & \ddots & \vdots \\ k_1(\mathbf{x}_1^n, \mathbf{x}_1^1)\mathbf{K}_2 & k_1(\mathbf{x}_1^n, \mathbf{x}_1^2)\mathbf{K}_2 & \cdots & k_1(\mathbf{x}_1^n, \mathbf{x}_1^n)\mathbf{K}_2 \end{pmatrix}, \tag{15}$$

where $\otimes$ indicates the Kronecker product and $\mathbf{x}_1^n$ refers to the n^{th} samples in data matrix $\mathbf{X}_1$. The pairwise kernel $\mathbf{K}$, obtained from Kronecker product, can substitute the $\mathbf{K}_x$ in the equation 13. The solution for pairwise KRR is given in the Equation 16.

$$\alpha = (\mathbf{K} + \lambda \mathbf{I}_N)^{-1} y. \tag{16}$$

It easily can be observed that the size of pairwise kernel $\mathbf{K}$ increases rapidly with the growth in the number of samples. Therefore, obtaining a model on the train set can be computationally infeasible regarding both memory and time usage. However, this problem can be overcome by using several algebraic properties of the Kronecker product. These computational short-cuts can be used to accelerate the training phase. Before explaining these short-cuts, several algebraic properties are presented first (Cichońska et al., 2018).

The vectorization operator, $vec(\cdot)$, stores the columns of a matrix into a vector. For instance, imagine matrix $\mathbf{Y} \in \mathbb{R}^{n \times m}$ then $vec(\mathbf{Y}) = \mathbf{y}$. Additionally, for to arbitrary

matrices $\mathbf{A}$ and $\mathbf{B}$, we have; $(\mathbf{A}\otimes\mathbf{B})vec(\mathbf{D}) = vec(\mathbf{B}\mathbf{D}\mathbf{A}^T)$, $(\mathbf{A}\otimes\mathbf{B})(\mathbf{D}\otimes\mathbf{E}) = (\mathbf{A}\mathbf{D})\otimes(\mathbf{B}\mathbf{E})$, and $(\mathbf{A}\otimes\mathbf{B})^{-1} = \mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$.

Now by knowing these properties, we explain the computational short-cut. Given two data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ and their Kronecker product $\mathbf{K}$, the eigenvalue decomposition of the Kronecker product matrix can be obtained from the eigenvalue decomposition of each data matrix. That is, if $\mathbf{K}_1 = \mathbf{U}_1\Lambda_1\mathbf{U}_1^{-1}$ and $\mathbf{K}_2 = \mathbf{U}_2\Lambda_2\mathbf{U}_2^{-1}$ are eigenvalue decompositions of both data matrices where $\mathbf{U}_1 \in \mathbb{R}^{n\times n}$ and $\mathbf{U}_2 \in \mathbb{R}^{m\times m}$ are orthogonal matrices comprise eigenvectors in their columns and $\mathbf{\Lambda}_1 \in \mathbb{R}^{n\times n}$ and $\mathbf{\Lambda}_2 \in \mathbb{R}^{m\times m}$ are diagonal matrices that store corresponding eigenvalue of $\mathbf{K}_1$ and $\mathbf{K}_2$ the pairwise KRR with the Kronecker product has a closed form and it is presented in the Equation 17.

$$\begin{aligned}
\boldsymbol{\alpha} &= (\mathbf{K}_1\otimes\mathbf{K}_2 + \lambda\mathbf{I}_N)^{-1}vec(\mathbf{Y}) \\
&= ((\mathbf{U}_1\mathbf{\Lambda}_1\mathbf{U}_1^{-1})\otimes(\mathbf{U}_2\mathbf{\Lambda}_2\mathbf{U}_2^{-1}) + \lambda\mathbf{I}_N)^{-1}vec(\mathbf{Y}) \\
&= ((\mathbf{U}_1\otimes\mathbf{U}_2)(\mathbf{\Lambda}_1\otimes\mathbf{\Lambda}_2 + \lambda\mathbf{I}_N)(\mathbf{U}_1^{-1}\otimes\mathbf{U}_2^{-1}))^{-1}vec(\mathbf{Y}) \\
&= (\mathbf{U}_1\otimes\mathbf{U}_2)(\mathbf{\Lambda}_1\otimes\mathbf{\Lambda}_2 + \lambda\mathbf{I}_N)^{-1}(\mathbf{U}_1^{-1}\otimes\mathbf{U}_2^{-1})vec(\mathbf{Y}) \\
&= (\mathbf{U}_1\otimes\mathbf{U}_2)(\mathbf{\Lambda}_1\otimes\mathbf{\Lambda}_2 + \lambda\mathbf{I}_N)^{-1}vec(\mathbf{U}_2^T\mathbf{Y}\mathbf{U}_1) \\
&= (\mathbf{U}_1\otimes\mathbf{U}_2)vec(\mathbf{R}) \\
&= vec(\mathbf{U}_2\mathbf{R}\mathbf{U}_1^T),
\end{aligned} \tag{17}$$

where $vec(\mathbf{R}) = (\mathbf{\Lambda}_1\otimes\mathbf{\Lambda}_2 + \lambda\mathbf{I}_N)^{-1}vec(\mathbf{U}_2^T\mathbf{Y}\mathbf{U}_1)$. Additionally, the Kronecker product of two diagonal matrices yields a diagonal matrix. Thus, $diag(\mathbf{\Lambda}_1\otimes\mathbf{\Lambda}_2) = diag(\mathbf{\Lambda}_1)\otimes diag(\mathbf{\Lambda}_2) = vec(diag(\mathbf{\Lambda}_2)diag(\mathbf{\Lambda}_2^T))$. By using this property we can prevent the explicit construction of any enormous Kronecker product matrix. Similar to the KRR, in the pairwise KRR the prediction for a unseen sample pair, $(\mathbf{z}_1, \mathbf{z}_2)$, can be obtain by using the Equation 18.

$$f(\mathbf{z}^1, \mathbf{z}^2) = \mathbf{k}^T\alpha = (\mathbf{k}_1^T\otimes\mathbf{k}_2^T)\alpha = (\mathbf{k}_1^T\otimes\mathbf{k}_2^T)vec(\mathbf{U}_2\mathbf{R}\mathbf{U}_1^T) = \mathbf{k}_2^T(\mathbf{U}_2\mathbf{R}\mathbf{U}_d^T)\mathbf{k}_1, \tag{18}$$

where the kernel values between unseen instances of the data matrix $\mathbf{X}^1$, $\mathbf{z}^1$, and each train samples of the same data matrix, $\mathbf{x}_1$, are stored in the vector $\mathbf{k}_1$, whereas, $\mathbf{k}_2$ is a vector comprises of kernel values between unseen samples of the data matrix $\mathbf{X}_2$, $\mathbf{z}^2$, and train samples of the data matrix $\mathbf{X}_2$, $\mathbf{x}^2$ (Cichońska et al., 2018; Pahikkala et al., 2013)

3.3 Canonical Correlation Analysis

In 1936, Hotelling introduced Canonical Correlation Analysis (i.e. CCA) to statistical study for two-view multivariate Hotelling (1936). Since then this method has been used in many fields such as economics, psychology, geography, medicine, chemistry, and etc. In the past decade, CCA has been used commonly in the modern field such as machine learning and data analysis, bioinformatics and computational biology, neuroscience. In these studies, CCA has been employed for dimensionality reduction (Uurtio et al., 2018b).

Canonical Correlation Analysis reviles the linear relationships between two variable sets, (Thompson, 2005). These variable sets (also called views) indicate different

measurements on a set of individuals (i.e. observation or samples), (Thompson, 2005; Uurtio et al., 2018b). One good example of different variable sets can be found in medical studies. Assume the information of several patients (i.e. samples) having a specific disease can be divided and store in two different variable sets;

1 Variables that describe the clinical symptoms of the disease,

2 Biological data (e.g gene expression, and mutation) extracted from tissue samples of patients.

These views, we call them view a and view b, are denoted by two row matrices $\mathbf{X}_a \in \mathbb{R}^{n\times p}$ and $\mathbf{X}_b \in \mathbb{R}^{n\times q}$, respectively. The samples are considered to be from a normal multivariate distribution. Additionally, features (variables) for n sample are presented by column vectors $\mathbf{a}_i \in \mathbb{R}^n$ for $i = 1,2,\cdots,p$ and $\mathbf{b}_j \in \mathbb{R}^n$ for $j = 1,2,\cdots,q$ (Uurtio et al., 2018b; Anderson et al., 1958).

The aim of canonical correlation analysis is to find a pair of linear transformation that maximizes the correlation between transformed variable sets. This transformation can be obtained by an inner product between a pair of weight vector and variables. These weight vectors are referred to as canonical weight vectors (in short canonical vectors) in literature. Variable sets $\mathbf{X}_a \in \mathbb{R}^{n\times p}$ and $\mathbf{X}_b \in \mathbb{R}^{n\times q}$ can be linearly transformed by canonical weight vectors $\mathbf{w}_a \in \mathbb{R}^p$ and $\mathbf{w}_b \in \mathbb{R}^q$, respectively as follow;

$$\langle \mathbf{X}_a, \mathbf{w}_a \rangle = \mathbf{z}_a \text{ and } \langle \mathbf{X}_b, \mathbf{w}_b \rangle = \mathbf{z}_b,$$

where $\mathbf{z}_a \in \mathbb{R}^n$ and $\mathbf{z}_b \in \mathbb{R}^n$, which are also called as canonical variables or canonical scores (Uurtio et al., 2018b; Anderson et al., 1958).

The CCA method endeavors to maximize the correlation between image vectors (i.e. canonical variables) $\mathbf{z}_a$ and $\mathbf{z}_b$ in the common subspace $\mathbb{R}^n$. That is, the cosine of the angle between images $\mathbf{z}_a$ and $\mathbf{z}_b$, which can be calculated by formula $cos(\mathbf{z}_a, \mathbf{z}_b) = \frac{<\mathbf{z}_a, \mathbf{z}_b>}{\|\mathbf{z}_a\|\|\mathbf{z}_b\|}$, is equivalent to the correlation (measure of closeness) between images. Therefore, the aim in CCA is to minimize the angle between two images $\mathbf{z}_a$ and $\mathbf{z}_b$. In order to obtain unique values for the canonical weight vectors, they are constrained to be unit norm vectors and their enclosing angle is considered to be $\theta \in [0, \frac{\pi}{2}]$. The cosine of the angle θ is referred to as the canonical correlation in the literature. Due to the unit norm constraint the former formula is equal to $cos(\mathbf{z}_a, \mathbf{z}_b) = \langle \mathbf{z}_a, \mathbf{z}_b \rangle$ (Uurtio et al., 2018b; Golub and Zha, 1995).

The first canonical correlation between the canonical variables (i.e. images) $\mathbf{z}_a$ and $\mathbf{z}_b$ of position vectors (i.e. canonical weight vector) $\mathbf{w}_a$ and $\mathbf{w}_b$ result in the highest correlation. That is, the smallest angle, θ_1, between images $\mathbf{z}_a$ and $\mathbf{z}_b$. The $cos\theta_1$ indicate the first canonical correlation. Therefore;

$$\begin{aligned} cos\theta_1 &= \max_{\mathbf{z}_a, \mathbf{z}_b \in \mathbb{R}^n} \langle \mathbf{z}_a, \mathbf{z}_b \rangle, \\ \|\mathbf{z}_a\|_2 = 1 \quad \|\mathbf{z}_b\|_2 &= 1. \end{aligned} \tag{19}$$

If we consider the $\mathbf{z}_a^1$ and $\mathbf{z}_b^1$ are the first pair of canonical variables that determine the maximum correlation, the second pair of canonical variables, $\mathbf{z}_a^2$ and $\mathbf{z}_b^2$, are orthogonal

to the first pair and has the second smallest enclosing angle θ_2. Thus, r finite number of pairs for canonical variables can be found recursively by removing all previous pairs from orthogonal complements set of canonical variables. All the r enclosing angles between pairs, $\theta_r \in [0, \frac{\pi}{2}]$ for $r = 1, 2, \cdots, q$ when $p > q$, are found in the ascending order and can be determined as presented in the Equation 20.

$$\begin{aligned} cos\theta_r &= \max_{\mathbf{z}_a, \mathbf{z}_b \in \mathbb{R}^n} \langle \mathbf{z}_a^r, \mathbf{z}_b^r \rangle, \\ s.t. \left\| \mathbf{z}_a^r \right\|_2 &= 1 \quad \left\| \mathbf{z}_b^r \right\|_2 = 1, \\ \langle \mathbf{z}_a^r, \mathbf{z}_a^j \rangle &= 0 \quad \langle \mathbf{z}_b^r, \mathbf{z}_b^j \rangle = 0, \\ \forall\, j \neq r : j, r &= 1, 2, \cdots, min(p, q). \end{aligned} \tag{20}$$

Additionally, it can be seen that the number of pairs of canonical variables, r, indicate the dimensionality of CCA, and consequently; is equal to the number of patterns that are extracted from the data. The importance of these found patterns, all the canonical weight vector $\mathbf{w}_a$ and $\mathbf{w}_b$ and all canonical variables (i.e. images) $\mathbf{z}_a$ and $\mathbf{z}_b$, is indicated by canonical correlation and they tend to decrease as the number of patterns increase (Uurtio et al., 2018b).

There are several methods introduced for obtaining the canonical weight vector $\mathbf{w}_a$ and $\mathbf{w}_b$ and respectively canonical variables $\mathbf{z}_a$ and $\mathbf{z}_b$. The original method, proposed by Hotelling (1936), includes solving a standard eigenvalue problem. In 1957, Healy proposed a singular value decomposition (SVD) based solution to find the positions and the images of the CCA method (Healy, 1957). In 2002 and 2004, a generalized eigenvalue problem has been suggested in the works of Bach and Jordan (2002); Hardoon et al. (2004) for obtaining the canonical variables. In the following, we explain the original method proposed by Hotelling, using standard eigenvalue method for solving CCA.

Standard Eigenvalue Technique for Solving CCA

Hotelling proposed the use of standard eigenvalue problem in order to obtain both pairs of canonical weights, $\mathbf{w}_a$ and $\mathbf{w}_b$, and pairs of canonical variables, $\mathbf{z}_a$ and $\mathbf{z}_b$. In his suggested solution, the characteristic equation is obtained by using the Lagrange multiplier. Let assume two data matrices $\mathbf{X}_a$ and $\mathbf{X}_b$ of sizes $n \times p$ and $n \times q$. Then the covariance matrix of samples, donated by $\mathbf{C}_{ab}$, is calculated by $\mathbf{C}_{ab} = \frac{1}{n-1}\mathbf{X}_a^T\mathbf{X}_b$ between variables of data matrices $\mathbf{X}_a$ and $\mathbf{X}_b$. Moreover, the empirical variance matrices for data matrices $\mathbf{X}_a$ and $\mathbf{X}_b$ are obtained by $\mathbf{C}_{aa} = \frac{1}{n-1}\mathbf{X}_a^T\mathbf{X}_a$ and $\mathbf{C}_{bb} = \frac{1}{n-1}\mathbf{X}_b^T\mathbf{X}_b$, respectively. The joint covariance matrix is constructed as given in the Equation 21.

$$\begin{pmatrix} \mathbf{C}_{aa} & \mathbf{C}_{ab} \\ \mathbf{C}_{ba} & \mathbf{C}_{bb} \end{pmatrix}. \tag{21}$$

The angle between the first pair of canonical variables (i.e. images) $\mathbf{z_a} = \mathbf{X}_a\mathbf{w_a}$ and $\mathbf{z_b} = \mathbf{X}_b\mathbf{w_b}$ is the smallest angle and consequently, it corresponds to the highest correlation. The correlation between a pair of canonical variables $\mathbf{z}_a$ and $\mathbf{z}_b$ is not depending on the scale of them, therefore; a constraint can be considered for canonical weight vector $\mathbf{w}_a$ and $\mathbf{w}_b$ in order to canonical variables $\mathbf{z}_a$ and $\mathbf{z}_b$ have unit variance

(Hotelling, 1936; Uurtio et al., 2018b). Therefore; the following can be obtained:

$$\mathbf{z}_a^T \mathbf{z}_a = \mathbf{w}_a^T \mathbf{X}_a^T \mathbf{X}_a \mathbf{w}_a = \mathbf{w}_a^T \mathbf{C}_{aa} \mathbf{w}_a = 1, \tag{22}$$

$$\mathbf{z}_b^T \mathbf{z}_b = \mathbf{w}_b^T \mathbf{X}_b^T \mathbf{X}_b \mathbf{w}_b = \mathbf{w}_b^T \mathbf{C}_{bb} \mathbf{w}_b = 1. \tag{23}$$

The variables of $\mathbf{X}_a$ and $\mathbf{X}_b$ are considered to be centered and have zero means, thus, they are considered to be normally distributed and comparable. Under this condition, the covariance between two canonical variables $\mathbf{z}_a$ and $\mathbf{z}_b$ can be obtained by the Equation 24:

$$\mathbf{z}_a^T \mathbf{z}_b = \mathbf{w}_a^T \mathbf{X}_a^T \mathbf{X}_b \mathbf{w}_b = \mathbf{w}_a^T \mathbf{C}_{ab} \mathbf{w}_b. \tag{24}$$

Accordingly, we can use Equations 22, 23, and 24 and substitute them in the Equation 19 to obtain the Equation 25.

$$\begin{gathered} \cos\theta = \max\langle \mathbf{z}_a, \mathbf{z}_b \rangle = \max\langle \mathbf{w}_a^T \mathbf{C}_{a,b} \mathbf{w}_b \rangle, \\ \|\mathbf{z}_a\|_2 = \sqrt{\mathbf{w}_a^T \mathbf{C}_{aa} \mathbf{w}_a} = 1 \quad \|\mathbf{z}_b\|_2 = \sqrt{\mathbf{w}_b^T \mathbf{C}_{bb} \mathbf{w}_b} = 1. \end{gathered} \tag{25}$$

The solution for the aforementioned problem can be acquired by using the Lagrange multiplier techniques. The Lagrange formulation of the Equation 25 is as the Equation 26.

$$L = \mathbf{w}_a^T \mathbf{C}_{ab} \mathbf{w}_b - \frac{\lambda_1}{2}(\mathbf{w}_a^T \mathbf{C}_{aa} \mathbf{w}_a - 1) - \frac{\lambda_2}{2}(\mathbf{w}_b^T \mathbf{C}_{bb} \mathbf{w}_b - 1), \tag{26}$$

where λ_1 and λ_2 refer to the Lagrange multipliers. To find the solution $\mathbf{w}_a$ and $\mathbf{w}_b$ for the Equation 26 we differentiating L with respect to $\mathbf{w}_a$ and $\mathbf{w}_b$. Additionally, we know that both $\mathbf{w}_a^T \mathbf{C}_{aa} \mathbf{w}_a = 1$ and $\mathbf{w}_b^T \mathbf{C}_{bb} \mathbf{w}_b = 1$, thus it easily can be proved that $\lambda_1 = \lambda_2 = \lambda$. Therefore, the solution for optimal $\mathbf{w}_a$ can be obtain by Equation 27.

$$\mathbf{w}_a = \frac{\mathbf{C}_{aa}^{-1} \mathbf{C}_{ab} \mathbf{w}_b}{\lambda}, \tag{27}$$

and consequently Equation 28 is obtained.

$$\frac{1}{\lambda} \mathbf{C}_{ba} \mathbf{C}_{aa}^{-1} \mathbf{C}_{ab} \mathbf{w}_b - \lambda \mathbf{C}_{bb} \mathbf{w}_b = 0, \tag{28}$$

by manipulating the Equation 28 we can obtain a generalized eigenvalue problem as in the Equation 29;

$$\mathbf{C}_{ba} \mathbf{C}_{aa}^{-1} \mathbf{C}_{ab} \mathbf{w}_b = \lambda^2 \mathbf{C}_{bb} \mathbf{w}_b. \tag{29}$$

The above equation can be reduced to a standard eigenvalue problem if $\mathbf{C}_{bb}$ is invertible. This equation is illustrated in the Equation 30;

$$\mathbf{C}_{bb}^{-1} \mathbf{C}_{ba} \mathbf{C}_{aa}^{-1} \mathbf{C}_{ab} \mathbf{w}_b = \lambda^2 \mathbf{w}_b. \tag{30}$$

Thus, the CCA can be solve by a standard eigenvalue solution of the matrix $\mathbf{C}_{bb}^{-1} \mathbf{C}_{ba} \mathbf{C}_{aa}^{-1} \mathbf{C}_{ab}$ have eigenvalues. These eigenvalues can be found by solving the Characteristic Equation 31;

$$|\mathbf{C}_{bb}^{-1} \mathbf{C}_{ba} \mathbf{C}_{aa}^{-1} \mathbf{C}_{ab} - \lambda^2 \mathbf{I}| = 0. \tag{31}$$

The canonical correlations are equal to the square roots of the acquired eigenvalues (Hotelling, 1936; Uurtio et al., 2018b).

3.3.1 Extending the CCA by Regularisation Technique

The CCA has been adopted in the studies that the number of observations is higher than the number of variables in both views. The singularity of variance matrices $\mathbf{C}_{aa}$ and $\mathbf{C}_{bb}$ can cause that these matrices cannot be inverted when we want to employ the standard eigenvalue technique to solve the CCA. Moreover, the square-root factors of singular variance matrices $\mathbf{C}_{aa}$ and $\mathbf{C}_{bb}$ my not exist when the SVD method has been adopted. If the number of observations exceeds the number of variables, the singularity of variance matrices possibly will overcome. On the other hand, if the number of observations is less than the number of variables the collinearity between the variables increases. For the first time, Vinod proposed a solution to tackle the collinearity problem in the studies with the insufficient sample size (Uurtio et al., 2018b). This technique is called regularisation and in the following, we give an overview of this popular technique.

In the work of Vinod (1976) the singularity problem is adopted, and it is proposed to add a regularisation term to the diagonal of both variance matrices. These regularisation values are arbitrary constant that improves the variance matrices' invertibility. In the Equation 32, the modified constraints of CCA that use the constructed variance matrices by additional regularized constant $c_1 \geq 0$ and $c_1 \geq 0$ is illustrated.

$$\begin{aligned} \mathbf{w}_a^T(\mathbf{C}_{aa} + c_1\mathbf{I})\mathbf{w}_a = 1, \\ \mathbf{w}_b^T(\mathbf{C}_{bb} + c_2\mathbf{I})\mathbf{w}_b = 1. \end{aligned} \tag{32}$$

In comparison with the original method, the magnitudes of the canonical weight vectors, $\mathbf{w}_a$ and $\mathbf{w}_b$, in the regularised CCA are smaller due to the additional regularisation term in the Equation 32. Hence, the constraints of the regularised CCA in the optimization problem is changed as it is presented in the Equation 33.

$$\begin{aligned} \cos\theta \quad &= \quad \max_{\mathbf{w}_a \in \mathbb{R}^p, \mathbf{w}_b \in \mathbb{R}^q} \quad (\mathbf{w}_a^T \mathbf{C}_{ab} \mathbf{w}_b), \\ \text{s.t} \quad &\mathbf{w}_a^T(\mathbf{C}_{aa} + c_1\mathbf{I})\mathbf{w}_a = 1 \quad \mathbf{w}_b^T(\mathbf{C}_{bb} + c_2\mathbf{I})\mathbf{w}_b = 1. \end{aligned} \tag{33}$$

The same as in the case of original CCA, the canonical weight vectors in the regularised CCA can be obtained by employing standard and generalized eigenvalues techniques Uurtio et al. (2018b). These two solutions are presented in the Equations 34 and 35, respectively.

$$(\mathbf{C}_{bb} + c_2\mathbf{I})^{-1}\mathbf{C}_{ba}(\mathbf{C}_{aa} + c_1\mathbf{I})^{-1}\mathbf{C}_{ab}\mathbf{w}_b = \lambda^2\mathbf{w}_b, \tag{34}$$

$$\begin{pmatrix} 0 & \mathbf{C}_{ab} \\ \mathbf{C}_{ba} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{C}_{aa} + c_1\mathbf{I} & 0 \\ 0 & \mathbf{C}_{bb} + c_2\mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{pmatrix}. \tag{35}$$

Similar to the CCA method, The canonical correlations of the regularised CCA is obtained by the inner product of each pair of canonical variables $< \mathbf{z}_a^i, \mathbf{z}_b^i >$, where $i = 1, 2, \cdots, min(p, q)$.

3.3.2 Introducing Sparsity to the the CCA Method

In both cases of CCA and regularised CCA, the entries of the canonical correlation weight vectors indicate the linear relation between the variables. These relations are interpretable when the number of variables is not high. Nevertheless, more often the number of variables in the modern datasets is much higher than it could be sufficient for the human to study and interpret them. To avoid such a problem, the coefficient of the canonical weight vectors should be constrained in a way that the only subset of them obtain non-zero value.

In order to enforce sparsity into the canonical weight vectors, the soft-thresholding operators have been used on the canonical weight vectors. Thus; a subset of variables will obtain non-zero weight and the linear relation between these entries of canonical weight vectors can be studied in the data space. This method has been proposed by Parkhomenko in 2007 for the first time and does not require prior knowledge over the variables of each view (Parkhomenko et al., 2007; Uurtio et al., 2018b).

The canonical weight vectors are obtained by using the SVD method as it has been presented in the original CCA method. The sparsity has been induced by sparse singular value decomposition of the matrix $\mathbf{C}_{aa}^{-\frac{1}{2}}\mathbf{C}_{ab}\mathbf{C}_{bb}^{-\frac{1}{2}}$. Accordingly, the left and right orthonormal singular vectors, $\mathbf{U}$ and $\mathbf{V}$, have sparse entries by applying the L_1 norm constrain on them. To find the sparse canonical weight vectors by using SVD technique the recursive Algorithm 1 has been proposed by Parkhomenko et al. (2007).

Algorithm 1: Sparse Canonical Correlation Analysis Algorithm

1 Initialization: Select the sparsity indicators c_a and c_b for both canonical weight vector. Initialize the canonical weight vectors $\mathbf{u}_0$ and $\mathbf{v}_0$, and set $i = 0$;
2 **repeat**
3 Update the left vector $\mathbf{u}$ as follows:
4 a- $\mathbf{u}_{i+1} = \mathbf{C}_{aa}^{-\frac{1}{2}}\mathbf{C}_{ab}\mathbf{C}_{bb}^{-\frac{1}{2}}\mathbf{v}_i$ and normalize it
5 b- Soft-thresholding: $\mathbf{u}_{i+1} = \left(\left\|\mathbf{u}_{i+1} - \frac{1}{2}c_a\right\|\right) + sign(\mathbf{u}_{i+1})$
6 c- Normalizing $\mathbf{u}_{i+1}$ again
7 Update the right vector $\mathbf{v}$ as follows:
8 a- $\mathbf{v}_{i+1} = \mathbf{C}_{aa}^{-\frac{1}{2}}\mathbf{C}_{ab}\mathbf{C}_{bb}^{-\frac{1}{2}}\mathbf{u}_{i+1}$ and normalized it
9 b- Soft-thresholding: $\mathbf{v}_{i+1} = \left(\left\|\mathbf{v}_{i+1} - \frac{1}{2}c_b\right\|\right) + sign(\mathbf{v}_{i+1})$
10 c- Normalizing $\mathbf{v}_{i+1}$ again
11 **until** *Until convergence*;
12 **return** $\mathbf{u}$, $\mathbf{v}$

The algorithm 1 returns the first canonical weight vectors pair. In order to obtain the consecutive pairs of canonical weight vectors, the matrix $\mathbf{C}_{aa}^{-\frac{1}{2}}\mathbf{C}_{ab}\mathbf{C}_{bb}^{-\frac{1}{2}}$ should be deflated and the new patterns can be found by using SVD techniques on the obtained matrix (Parkhomenko et al., 2007; Uurtio et al., 2018b).

3.3.3 Kernel Canonical Correlation Analysis

The CCA methods originally is used for finding the linear relation between variables of two views in the data space, $\mathbf{X}_a \in \mathbb{R}^{n\times p}$ and $\mathbf{X}_b \in \mathbb{R}^{n\times q}$ where n is the number of samples, and p, q correspond to the number of variables in view a and b, respectively. The images (i.e. canonical variables) $\mathbf{z}_a$ and $\mathbf{z}_b$ are obtained by inner product of position vectors (canonical weight vectors) $\mathbf{w}_a \in \mathbb{R}^p$ and $\mathbf{w}_b \in \mathbb{R}^q$ and corresponding views, $\mathbf{z}_a = \mathbf{X}_a\mathbf{w}_a$ and $\mathbf{z}_b = \mathbf{X}_b\mathbf{w}_b$. These canonical weight vectors are obtained such as the enclosing angle between their corresponding image is minimized and they indicate the relations between the features in each view. Both canonical weigh vectors $\mathbf{w}_a$ and $\mathbf{w}_b$ and their corresponding canonical variables $\mathbf{z}_a$ and $\mathbf{z}_b$ are acquired in the data space which is Euclidean space; therefore, only the linear relation can be extracted (Uurtio et al., 2018b).

The non-linear relations between features can be extracted when the canonical weight vectors $\mathbf{w}_a$ and $\mathbf{w}_b$ are obtained in non-linear feature space. To this end, the kernel method can be used. Firstly, the original samples, $\mathbf{X}_a^i$ and $\mathbf{X}_b^i$ where $i = 1, 2, \cdots, n$ correspond to the sample size, are mapped into non-linear Hilbert space $\mathscr{H}_a$ and $\mathscr{H}_b$. These feature maps are presented in the equation 36.

$$\begin{aligned} \phi_a : \mathbb{R}^p &\mapsto \mathscr{H}_a \\ \phi_b : \mathbb{R}^q &\mapsto \mathscr{H}_b. \end{aligned} \tag{36}$$

Then, symmetric positive semi-definite kernels are employed in order to capture the similarity between samples through an inner product of obtained high dimensional Hilbert (i.e. feature) spaces. The Equation 37 demonstrate these kernels.

$$\begin{aligned} \mathbf{K}_a(\mathbf{x}_a^i, \mathbf{x}_a^j) &= \langle \phi_a(\mathbf{x}_a^i), \phi_a(\mathbf{x}_a^j) \rangle_{\mathscr{H}_a} \\ \mathbf{K}_b(\mathbf{x}_b^i, \mathbf{x}_b^j) &= \langle \phi_b(\mathbf{x}_b^i), \phi_b(\mathbf{x_b}^j) \rangle_{\mathscr{H}_b}. \end{aligned} \tag{37}$$

These feature maps are non-linear; therefore, by substituting these kernels obtained from non-linear feature maps, instead of covariance matrices in Equations 22 and 23, the non-linear correlation between variables can be extracted (Bach and Jordan, 2002; Hardoon et al., 2004; Uurtio et al., 2018b). The kernel CCA is illustrated in the Figure 4.

The same as the original CCA, the canonical correlation vectors, $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^n$, are obtained in the feature space in such a way that minimizes the enclosing angle between canonical variable, $\mathbf{z}_a$ and $\mathbf{z}_b$ on the unit ball in $\mathbb{R}^n$ (Uurtio et al., 2018b; Hardoon et al., 2004). Therefore, the kernel CCA can be expressed as in the Equation 38.

$$\begin{aligned} cos(\mathbf{z}_a, \mathbf{z}_b) = \max_{\mathbf{z}_a, \mathbf{z}_b \in \mathbb{R}^n} \langle \mathbf{z}_a, \mathbf{z}_b \rangle &= \boldsymbol{\alpha}^T \mathbf{K}_a^T \mathbf{K}_b \boldsymbol{\beta}, \\ \text{s.t } \|\mathbf{z}_a\|_2 = \sqrt{\boldsymbol{\alpha}^T \mathbf{K}_a^2 \boldsymbol{\alpha}} &= 1 \\ \|\mathbf{z}_b\|_2 = \sqrt{\boldsymbol{\beta}^T \mathbf{K}_b^2 \boldsymbol{\beta}} &= 1, \end{aligned} \tag{38}$$

where the constraints are applied to guarantee the uniqueness of obtained canonical weight vectors. The Lagrange multiplier method can be applied to solve the optimization

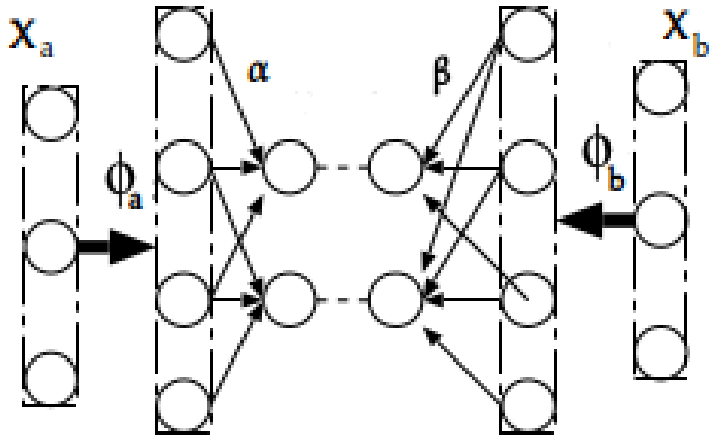


Figure 4: Kernel canonical correlation; $\mathbf{X}_a$ and $\mathbf{X}_b$ indicate the data matrices in view a and b respectively. ϕ_a and ϕ_b are feature maps for each view that map data points the into non-linear Hilbert spaces $\mathscr{H}_a$ and $\mathscr{H}_b$, respectively. α and β are the canonical weight vectors obtain in the feature space (AKAHO, 2001).

problem. This optimization problem is given in the Equation 39.

$$L = \boldsymbol{\alpha}^T \mathbf{K}_a^T \mathbf{K}_b \boldsymbol{\beta} - \frac{\lambda_1}{2}(\boldsymbol{\alpha}^T \mathbf{K}_a^2 \boldsymbol{\alpha} - 1) - \frac{\lambda_2}{2}(\boldsymbol{\beta}^T \mathbf{K}_b^2 \boldsymbol{\beta} - 1), \tag{39}$$

where λ_1 and λ_2 are the Lagrange multipliers. In order to acquire an optimize solution for canonical weight vectors, we drive the Equation 39 with respect to α and β. These derivation are shown in the Equations 40 and 41.

$$\frac{\delta L}{\delta \boldsymbol{\alpha}} = \mathbf{K}_a \mathbf{K}_b \boldsymbol{\beta} - \lambda_1 \mathbf{K}_a^2 \boldsymbol{\alpha} = 0, \tag{40}$$

$$\frac{\delta L}{\delta \boldsymbol{\beta}} = \mathbf{K}_b \mathbf{K}_a \boldsymbol{\alpha} - \lambda_2 \mathbf{K}_b^2 \boldsymbol{\beta} = 0. \tag{41}$$

By Multiplying the α^T into the Equation 40 from the left side and β^T into the Equation 41 from the left side result in the Equations 42 and 43.

$$\boldsymbol{\alpha}^T \mathbf{K}_a \mathbf{K}_b \boldsymbol{\beta} - \lambda_1 \boldsymbol{\alpha}^T \mathbf{K}_a^2 \boldsymbol{\alpha} = 0, \tag{42}$$

$$\boldsymbol{\beta}^T \mathbf{K}_b \mathbf{K}_a \boldsymbol{\alpha} - \lambda_2 \boldsymbol{\beta}^T \mathbf{K}_b^2 \boldsymbol{\beta} = 0. \tag{43}$$

From the Equation 38 we know that $\boldsymbol{\alpha}^T \mathbf{K}_a^2 \boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta}^T \mathbf{K}_b^2 \boldsymbol{\beta} = 1$; therefore, the result $\lambda_1 = \lambda_2 = \lambda$ is obtained and by substituting it in the Equation 40, the optimized $\boldsymbol{\alpha}$ determined by Equation 44.

$$\boldsymbol{\alpha} = \frac{\mathbf{K}_a^{-1} \mathbf{K}_a^{-1} \mathbf{K}_a \mathbf{K}_b \boldsymbol{\beta}}{\lambda} = \frac{\mathbf{K}_a^{-1} \mathbf{K}_b \boldsymbol{\beta}}{\lambda}. \tag{44}$$

Additionally, by replacing the alpha in the Equation 41 by $\boldsymbol{\alpha}$ obtained by the Equation 44, the solution for the optimize $\boldsymbol{\beta}$ is calculated as in the Equation 45.

$$\frac{1}{\lambda}\mathbf{K}_b\mathbf{K}_a\mathbf{K}_a^{-1}\mathbf{K}_b\boldsymbol{\beta} - \lambda\mathbf{K}_b^2\boldsymbol{\beta} = 0. \tag{45}$$

As it can be seen in the Equation 46, the generalized eigenvalue problem is formed by removing $\mathbf{K}_a\mathbf{K}_a^{-1}$.

$$\mathbf{K}_b^2\boldsymbol{\beta} = \lambda^2\mathbf{K}_b^2\boldsymbol{\beta}. \tag{46}$$

The generalized eigenvalue problem Equation 46 can simplified to a standard eigenvalue problem as in the Equation 47, when the kernel $\mathbf{K}_b^2$ is invertible.

$$\mathbf{I}\boldsymbol{\beta} = \lambda^2\boldsymbol{\beta}, \tag{47}$$

where the canonical correlation is stored in λ and $\boldsymbol{\beta}$ correspond to the canonical weight vectors in the feature space. In the feature space, all the Gram matrices are invertible. This means that $\lambda = 1$ for all canonical weight vectors, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and perfect correlation can be found (Hardoon et al., 2004; Uurtio et al., 2018b). This is because of the high-dimensionality of the feature space that is provided by kernel methods. Therefore, no meaningful result can be obtained from such naive application and the obtained non-linear relations between features are not explaining the input data sets under investigation. This is due to the over-fitting problem that occurs by high-dimension feature spaces (Hardoon et al., 2004).

In order to overcome the over-fitting problem in the kernel CCA method, regularisation technique is applied. This regularized parameter is applied on the norms of canonical weights vectors, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, to control the flexibility of projection through penalized factor (Hardoon et al., 2004; Bach and Jordan, 2002). The Equation 48 clarify the new constraint of canonical weight vector with additional regularised constant.

$$\begin{aligned} \alpha^T(\mathbf{K}_a + c_1\mathbf{I})^2\alpha = 1, \\ \beta^T(\mathbf{K}_b + c_2\mathbf{I})^2\beta = 1. \end{aligned} \tag{48}$$

As it can be seen the scalar regularisation values are added to the diagonal of the Gram matrices, $\mathbf{K}_a$ and $\mathbf{K}_b$ (Uurtio et al., 2018b). The result of the regularized kernel CCA can be obtained through the standard eigenvalue problem given in the Equation 49.

$$(\mathbf{K}_b + c_1\mathbf{I})^{-2}\mathbf{K}_b\mathbf{K}_a(\mathbf{K}_b + c_2\mathbf{I})^{-2}\mathbf{K}_a\mathbf{K}_b\alpha = \lambda^2\alpha. \tag{49}$$

Additionally, the same as the original CCA method, the generalized eigenvalue method can be employed in order to solve the kernel CCA. This can be obtained by replacing the data matrices, $\mathbf{X}_a$ and $\mathbf{X}_b$, with their corresponding Gram matrices, $\mathbf{K}_a$ and $\mathbf{K}_b$ (Hardoon et al., 2004; Uurtio et al., 2018b; Bach and Jordan, 2002). The solution of the generalized eigenvalue problem is presented in the Equation 50.

$$\begin{pmatrix} 0 & \mathbf{K}_a\mathbf{K}_b \\ \mathbf{K}_b\mathbf{K}_a & 0 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \lambda\begin{pmatrix} (\mathbf{K}_a + c_1\mathbf{I})^2 & 0 \\ 0 & (\mathbf{K}_b + c_2\mathbf{I})^2 \end{pmatrix} \tag{50}$$

The projection vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are obtained in the high-dimensional Reproducing Kernel Hilbert Space (i.e feature spaces) $\mathscr{H}_a$ and $\mathscr{H}_b$; therefore, they are not interpretable in the input space. That is, investigating the relation between original features is not possible in the kernel CCA. Additionally, kernel CCA considers all the variable in both views $\mathbf{X}_a$ and $\mathbf{X}_b$ and does not remove the irrelevant features and eventually affect the robustness of kernel CCA. The careful selection of the regularization parameter can overcome the latter problem (Chang et al., 2013).

3.3.4 Sparse Canonical Correlation Analysis through Hilbert-Schmidt Independence Criterion Optimization

In this section, we go through an extension of CCA that allows us;

1) Find linear and non-linear relations among variables of two views while inducing sparsity.

2) Obtain multiple canonical weight vectors that are interpretable (Chang et al., 2013; Uurtio et al., 2018a).

This sparse variation of CCA (i.e. HSIC-SCCA) use Hilbert-Schmidt Independence Criterion in order to find (in)dependency among variables of each view. Firstly, we go through some definition that is essential for understanding HSIC-SCCA method.

Definition 3.9. Hilbert-Schmidt Norm: The Hilbert-Schmidt (HS) norm of a linear operator C (i.e. $C : \mathscr{G} \rightarrow \mathscr{F}$) can be calculated through Equation 51.

$$\|C\|_{HS}^2 := \sum_{i,j} \langle C v_j, u_i \rangle_{\mathscr{F}}^2, \tag{51}$$

u_i and v_j indicate orthonormal bases for two RKHS $\mathscr{F}$ and $\mathscr{G}$ respectively (Gretton et al., 2005a).

Definition 3.10. Cross-Covariance Operator: Assume two topological separable spaces, $\mathscr{X}$ and $\mathscr{Y}$, and two Borel sets, Γ and Λ, such as $(\mathscr{X}, \Lambda)$ and $(\mathscr{Y}, \Gamma)$ that have probability measurements p_x and p_y, respectively. A linear operator $C_{xy} : \mathscr{G} \rightarrow \mathscr{F}$ is a cross-covariance operator on $(\mathscr{X} \times \mathscr{Y}, \Gamma \times \Lambda)$ with joint probability measurement $p_{x,y}$ and it can be obtain as in the Equation 52.

$$C_{xy} := \mathbf{E}_{x,y}[(\phi(x) - \mu_x) \otimes (\psi(y) - \mu_y)] = \underbrace{\mathbf{E}_{x,y}[\phi(x) \otimes \psi(y)]}_{:=\hat{C}_{xy}} - \underbrace{\mu_x \otimes \mu_y}_{:=M_{xy}}, \tag{52}$$

where the mean elements, μ_x and μ_y, can be obtained by using the Equation 53 when $\mathscr{F}$ and $\mathscr{G}$ are two RKHS, and $\boldsymbol{f} \in \mathscr{F}$ and $\boldsymbol{g} \in \mathscr{G}$.

$$\begin{aligned} \langle \mu_x, \boldsymbol{f} \rangle_{\mathscr{F}} &:= \mathbf{E}_x[\langle \phi(x), \boldsymbol{f} \rangle_{\mathscr{F}}] = \mathbf{E}_x[f(x)], \\ \langle \mu_y, \boldsymbol{g} \rangle_{\mathscr{G}} &:= \mathbf{E}_y[\langle \psi(y), \boldsymbol{g} \rangle_{\mathscr{G}}] = \mathbf{E}_y[g(y)] \end{aligned} \tag{53}$$

In the aforementioned equation, ϕ and χ are feature maps from $\mathscr{X}$ and $\mathscr{Y}$ to $\mathscr{F}$ and $\mathscr{G}$ respectively (Gretton et al., 2005a).

Definition 3.11. Hilbert-Schmidt Independence Criterion: The Hilbert-Schmidt Independence Criterion for two given separable RKHS $\mathscr{F}$ and $\mathscr{G}$ and a joint probability measurement p_{xy} on $(\mathscr{X}\times\mathscr{Y},\Gamma\times\Lambda)$ can be calculated with the Equation 54.

$$HSIC(p_{xy},\mathscr{F},\mathscr{G}) := \|C_{xy}\|^2_{HS}, \tag{54}$$

where C_{xy} refers to cross covariance operator. The Equation 54 can be obtain through kernel functions as presented in the Equation 55.

$$\begin{aligned} HSIC(p_{xy},\mathscr{F},\mathscr{G}) = {} & \mathbf{E}_{x^i,x^j,y^i,y^j}[k_a(x^i,x^j)l(y^i,y^j)] + \mathbf{E}_{x^i,x^j}[k(x^i,x^j)]\mathbf{E}_{y^i,y^j}[l(y^i,y^j)] \\ & - 2\mathbf{E}_{x^i,y^i}[\mathbf{E}_{x^j}[k(x^i,x^j)]\mathbf{E}_{y^j}[l(y^i,y^j)]], \end{aligned} \tag{55}$$

where k and l are bounded kernels (i.e. $\mathbf{E}_{x^i,x^j}[k(x^i,x^j)]<\infty$ and $\mathbf{E}_{y^i,y^j}[l(y^i,y^j)]<\infty$), and $\mathbf{E}_{x^i,x^j,y^i,y^j}$ indicates the expectation value of the independent pairs (x^i,y^i) and (x^j,y^j) obtain from p_{xy} (Gretton et al., 2005a).

The approximation of $HSIC(p_{xy},\mathscr{F},\mathscr{G})$ over a finite number of independent observations, n, in the two RKHS $\mathscr{F}$ and $\mathscr{G}$ can be obtain with the Equation 56.

$$HSIC(Z,\mathscr{F},\mathscr{G}) := \frac{trace(\hat{K}\hat{L})}{(n-1)^2}, \tag{56}$$

where Z refers to a set of independent observations drawn from p_{xy}, $Z := \{(x^1,y^1),(x^2,y^2),\cdots,(x^n,y^n)\}\subseteq\mathscr{X}\times\mathscr{Y}$, and $\hat{K}\in\mathbb{R}^{n\times n}$ and $\hat{L}\in\mathbb{R}^{n\times n}$ are centred Gram matrices, in the Hilbert spaces $\mathscr{F}$ and $\mathscr{G}$ Additionally, the HSIC can be calculated empirically by centring only one of the Gram matrices (Gretton et al., 2005b,a; Uurtio et al., 2018a). The HSIC is an indicator of (in)dependency for two universal kernel ($\mathscr{H}_a$ and $\mathscr{H}_b$) and can be apply in non-linear CCA, (Gretton et al., 2005a; Uurtio et al., 2018a).

From Section 3.3.2 of the previous chapter, we know that the sparsity can be inducted into the CCA by applying l_1-norm on the canonical weight vectors in the constraint of CCA. The level of this sparsity can be control through the l_1-norm constants c_1 and c_2 for each view a and b and a sparse subset of variable is selected by canonical weight vectors $\mathbf{u}\in\mathbb{R}^p$ and $\mathbf{v}\in\mathbb{R}^q$.

Additionally, non-linear relation between these variables can be obtained by using Gram matrices $\mathbf{K^u}$ and $\mathbf{K^v}$ where $\mathbf{K^u}=\langle\mathbf{u}^T\mathbf{x}_a^i,\mathbf{u}^T\mathbf{x}_a^j\rangle$ and $\mathbf{K^v}=\langle\mathbf{v}^T\mathbf{x}_b^i,\mathbf{v}^T\mathbf{x}_b^j\rangle$, and i and j indicate two different samples in views a and b. Thus, the HSIC-SCCA model is calculated by optimizing Equation 57

$$\begin{aligned} \max_{s.t.} \quad & \rho(\mathbf{u},\mathbf{v}) = \frac{trace(\hat{\mathbf{K}}^{\mathbf{u}}\hat{\mathbf{K}}^{\mathbf{v}})}{(n-1)^2} = \frac{\sum_{ij}\sum_{pq}k^{\mathbf{u}}_{ij}k^{\mathbf{v}}_{pq}}{(n-1)^2}, \\ & \|\mathbf{u}\|_1 \le c_1 \\ & \|\mathbf{v}\|_1 \le c_2 \end{aligned} \tag{57}$$

where n indicate the number samples, $k^{\mathbf{u}}_{ij}$ and $k^{\mathbf{v}}_{pq}$ refer to entries of centered kernels obtained from projected samples of views a and b respectively. Additionally, c_1 and c_2

refer to l_1-norm constant and control the level of sparsity (Uurtio et al., 2018a). It is possible to use the l_2-norm regularization instead of the l_1-norm in the situation that sparse subset of variables is not required. That is, the l_2-norm can easily substitute the l_1-norm. This method is called HSIC-CCA and has been proposed in the (Chang et al., 2013). Additionally, the l_1-norm can be applied to one of the views while the l_2-norm is used for the other one. By using l_2-norm the non-sparse projection vector for features is obtained and it is beneficial for the cases that sparse projection is not demanded (Uurtio et al., 2018a).

In order to obtain several canonical correlations, a deflation approach has to be taken. Deflation refers to the operation of removing the influence of an eigenvector by setting the corresponding eigenvalue of the eigenvector to zero. The Schur complement deflation techniques have been used in order to preserve the positive semidefinite of the data matrix (Mackey, 2009). Assume the m-th canonical weight vector, $\mathbf{u}^m$ and $\mathbf{v}^m$, for m-th canonical correlation component is calculated for data matrices $\mathbf{X_a}$ and $\mathbf{X_b}$. The $m+1$-th orthogonal canonical weight vectors, $\mathbf{u}^{(m+1)}$ and $\mathbf{v}^{(m+1)}$, can be obtained by applying Equation 57 to the data matrices obtain in Equation 58.

$$\begin{aligned} \mathbf{X}_a^{(m+1)} &= \mathbf{X}_a^{(m)} - \frac{\mathbf{u}^{(m)}\mathbf{u}^{(m)^T}\mathbf{X}_a^{(m)}}{\mathbf{u}^{(m)}\mathbf{u}^{(m)^T}}, \\ \mathbf{X}_b^{(m+1)} &= \mathbf{X}_b^{(m)} - \frac{\mathbf{v}^{(m)}\mathbf{v}^{(m)^T}\mathbf{X}_b^{(m)}}{\mathbf{v}^{(m)}\mathbf{v}^{(m)^T}}, \end{aligned} \tag{58}$$

where $\mathbf{X}_a^{(m)}$ and $\mathbf{X}_b^{(m)}$ are m-th data matrices in view a and b. This step guarantee that all calculated canonical weight vectors, $(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \cdots)$ and $(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \cdots)$, are mutually orthogonal (Uurtio et al., 2018a).

A recursive algorithm is suggested in (Uurtio et al., 2018a) to extract the m different non-linear relations between features of two views $\mathbf{X}_a$ and $\mathbf{X}_b$. This algorithm is presented in Algorithm 2.

Algorithm 2: Sparse Non-linear Canonical Correlation Analysis Algorithm

Inputs: $\mathbf{X}_a, \mathbf{X}_b, \epsilon$(convergence limit),
M (number of component),
R (number of repetition),
p_a (l_1 or l_2 norm type for $\mathbf{u}$),
p_b (l_1 or l_2 norm type for $\mathbf{u}$),
c_1 (regularization parameter for view a),
c_2 (regularization parameter for view b),
δ_u (parameters for $\mathbf{K}^u$),
δ_v (parameters for $\mathbf{K}^v$)
Outputs: $\mathbf{u}$ and $\mathbf{v}$
for m=1,2,$\cdots$,M **do**
  for r=1,2,$\cdots$,R **do**
    Initialize $\mathbf{u}_{mr}$ and $\mathbf{v}_{mr}$
    Compute $\mathbf{K}^u$, $\mathbf{K}^v$ and $\hat{\mathbf{K}}^v$
    repeat
      Compute $f_{old} = \rho(\mathbf{u},\mathbf{v})$
      Compute $\nabla_{\mathbf{u}} = \frac{\delta\rho(\mathbf{u},\mathbf{v})}{\delta\mathbf{u}}$
      Update $\mathbf{u}_{mr} = \Pi_{\|\cdot\|_{p_x}\leq c_1}(\mathbf{u}_{mr} + \gamma\nabla_{\mathbf{u}})$(line search to determine the stop size γ along the gradient)
      Compute $\mathbf{K}^u$, $\hat{\mathbf{K}}^u$
      Compute $\nabla_{\mathbf{v}} = \frac{\delta\rho(\mathbf{u},\mathbf{v})}{\delta\mathbf{v}}$
      Update $\mathbf{v}_{mr} = \Pi_{\|\cdot\|_{p_y}\leq c_2}(\mathbf{v}_{mr} + \gamma\nabla_{\mathbf{v}})$(line search to determine the stop size γ along the gradient)
      Compute $\mathbf{K}^u$, $\hat{\mathbf{K}}^u$
      Compute $f_{current} = \rho(\mathbf{u},\mathbf{v})$
    until $|f_{old} - f_{current}| < \epsilon$;
    $f_r = f_{current}$, $\mathbf{u}_r = \mathbf{u}_m r$, $\mathbf{v}_r = \mathbf{v}_m r$
  end
  Select $r^* = argmax f_r$,
  Store $\mathbf{U}(:,m) = \mathbf{u}_{r^*}$, $\mathbf{V}(:,m) = \mathbf{v}_{r^*}$
  Deflate $\mathbf{X}_a^{(m)}$, $\mathbf{X}_b^{(m)}$ by $\mathbf{U}(:,m)$ and $\mathbf{V}(:,m)$
end
return $\mathbf{u}$, $\mathbf{v}$

As it can be seen in the Algorithm 2, first the type of regularization technique, $l1-norm$ or $l2-norm$, and their values, c_1, c_2 are specified. Additionally, the number of components that are desired to obtained and the number of repetition has to be specified. In the sparse canonical correlation analysis algorithm, the projection vectors of each view are initialized randomly. The number of repetition indicates the number of random initialization of the projection vectors. The hyper-parameters for the kernels of each view are specified by δ_u and δ_v. The convergence limit, ϵ, is the stopping criteria of the algorithm. That is, for each component the optimization continues until the difference between HSIC measurements of the previous step and the current step is

smaller than ϵ. In order to optimize for both projection vector, $\mathbf{u}$ and $\mathbf{v}$, a stochastic gradient has been applied. That is, in each iteration a mini-batch of instances, i and j, has been selected randomly and the gradient of kernel function with respect to projection vector $\mathbf{u}$ and $\mathbf{v}$, $\nabla_{\mathbf{u}}$ and $\nabla_{\mathbf{v}}$, has been calculated with respect to the selected instances. The stochastic gradient has been proposed since the calculation for the exact gradient is too costly, $\mathcal{O}(n^2 d)$ (Uurtio et al., 2018a).

3.4 Model Selection And Model Evaluation

In order to model selection and model evaluation, a different variation of cross-validation (e.g. CV) can be used. In this technique, the data set under study is divided into two parts; one for the model selection, training data set, and the other for evaluation, test data set. In the model selection step, the model parameters are selected in order to the predictive model to obtain the best performance on the train data set. The accuracy of this optimal model is evaluated on the test data set which has not been seen before by the model in the training step (Krstajic et al., 2014).

3.4.1 K-Fold Cross-Validation

In the K-fold cross-validation technique, the dataset is partitioned into k equally sized segments. Additionally, accuracy criteria, such as the Pearson correlation, and an initial range of values for each model parameter is considered. Then an irritate algorithm obtain the accuracy of the model for all parameters in the given range. At each iteration, one of the partitions is used as a test set and $K-1$ folds are used for training the model. After obtaining the accuracy of the optimal model on the all iteration of test folds, the average prediction performance is reported as total model accuracy, (Krstajic et al., 2014).

3.4.2 Nested Cross-Validation

IIn the nested cross-validation technique, two-layer cross-validation is employed to select and evaluate a model. Same as the k-fold cross-validation, the data set is partitioned into the k different segments with equal size. These partitions are considered as the outer loop. At each iteration of the outer loop, one partition is considered for performance evaluation (test) and the rest folds are for model selection (train). Then the train set divided to n partition of equal size. These are inner loops which in them all combination of parameters are tested in the same as k-fold cross validation manner. The best parameters are selected based on the best performance of inner loops. These parameters are used in the training the model by an outer loop train set and tested on the test set of the outer loop. Similarly to the K-fold cross-validation, the accuracy of the model is reported by averaging over the accuracy of the result for outer loops (Krstajic et al., 2014).

4 Materials and Methods

In this chapter, first we go through the data sets that have been used in this study, Genomics of Drug Sensitivity in Cancer (e.g. GDSC) data set. Two methods, KCCA and HSIC-SCCA, are used for selecting features from cell line and drug compound data matrices of the given data set. The power of these methods, in selecting more relevant features, are evaluated by the accuracy of the model constructed by these features in prediction of drug sensitivity values. The work-flow of these experiments setups are presented fully in this chapter. Firstly, we introduce the data set that has been used in this study. After introducing the data set, the pre-processing step for the data set has been explained. At last, the setup of the experiments is presented.

Through this Chapter, the drug sensitivity data matrix is referred as $\mathbf{Y} \in \mathbb{R}^{n \times m}$ where rows are referring to drug compounds samples, whereas, cell lines samples are stored in columns. The cell lines data matrix, that comprises cell lines samples (rows) and gene expression values as features (columns) is indicated by $\mathbf{X}_c \in \mathbb{R}^{m \times q}$. The drug compounds data matrix, that store the drug compounds samples (rows) expressed by their fingerprints (columns), is indicated by $\mathbf{X}_D \in \mathbb{R}^{n \times p}$. Additionally, The drugs sensitivity responses kernel, cell lines kernel, and drug compounds kernel are indicated by $\mathbf{K}_R \in \mathscr{H}_R$, $\mathbf{K}_C \in \mathscr{H}_C$, and $\mathbf{K}_D \in \mathscr{H}_D$ and calculated by the Equation 59.

$$\begin{aligned} \mathbf{K}_R &= \langle \phi_R(\mathbf{y}^i), \phi_R(\mathbf{Y}^j) \rangle, \\ \mathbf{K}_C &= \langle \phi_C(\mathbf{x}_c^i), \phi_C(\mathbf{x}_c^j) \rangle, \\ \mathbf{K}_D &= \langle \phi_D(\mathbf{x}_d^i), \phi_D(\mathbf{x}_d^j) \rangle, \end{aligned} \tag{59}$$

where $\phi_R(\cdot), \phi_C(\cdot), \phi_D(\cdot)$ indicate the non-linear maps from the drug sensitivities input space, cell lines input space, and drug compounds input space to a non-linear feature space, respectively.

4.1 Genomics of Drug Sensitivity in Cancer (GDSC) Data set;

A subset of the large data set, Genomics of Drug Sensitivity in Cancer (i.e. GDSC), has been used in this study. Genomics of Drug Sensitivity in Cancer project has been initiated by Welcome Sanger Institute. This project has provided one of the largest publicly available data set for drug responses on the cancer cell lines since it has been established. GDSC contained the cancer drug sensitivity results of 73169 experiments only on its second released version (July 2012) that includes drug responses of 138 anticancer drugs on the range 329-668 cancer cell lines per a drug. This data set has been collected in order to facilitate the understanding of molecular and genomic changes under influence of drugs in the cancer research field, and it has been accompanied with detailed genomic data sets of cell lines (Yang et al., 2012).

4.1.1 Drug Sensitivity Data Matrix

A subset of GDSC data set has been selected for investigating the power of kernel CCA and HSIC-SCCA methods to extract the important features among the cell lines data

set, and drug compounds data set. That is the gene expression and fingerprints that are highly correlated with drug sensitivity values in non-linear space. Our data set is a subset comprises of 124 cell lines and 124 drug components that the drug responses are fully calculated. These drug-cell line responses are measured by fitting the dose-response curve through 9 points drug concentration measurements observed during 72 hours of drug treatment for each cell lines. These values are summarized as the natural log of IC_{50} value (micromolar concentration of a drug that inhibits half of the cell growth) (Ammad-ud din et al., 2016; Yang et al., 2012). The heat-map of this data matrix is illustrated in the Figure 5.

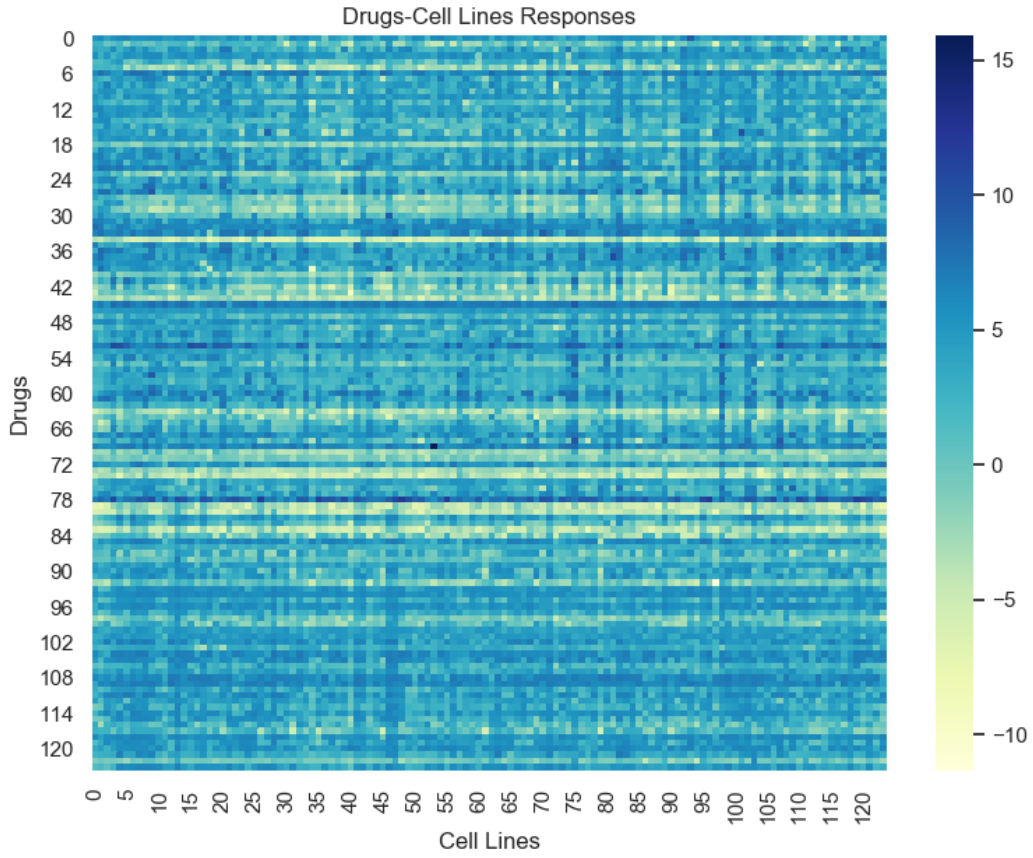


Figure 5: The heat-map of Drugs Responses data matrix. This data matrix contains the responses of 124 cell lines (columns) to 124 drugs (rows).

4.1.2 Cell Line-Gene Expression Data Matrix

Additional to the drug sensitivity data matrix, which is used for both feature selection task as input and later as the evaluation criteria, we require the genomic information of the cell lines. Therefore, the values of baseline gene expression measurements for all of the cell lines are used to capture genomic properties of cell lines. This data set contains

13321 gene expression values (Ammad-ud din et al., 2016). Figure 6 illustrate the gene expression data matrix for cell lines.

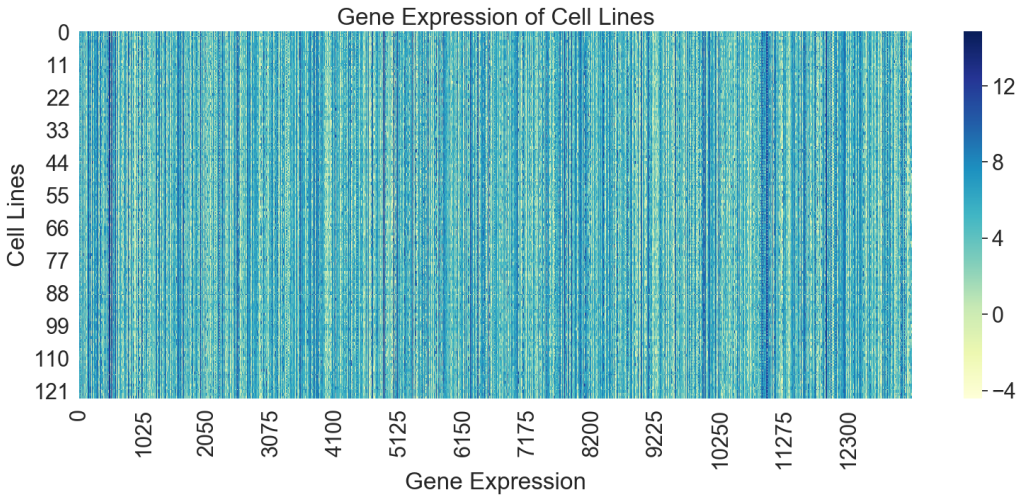


Figure 6: The heat-map of gene expression of cell lines. This data matrix contains the responses of 124 cell lines (columns) to 124 drugs (rows).

4.1.3 Drug Compound-Fingerprints Data Matrix

In this study, a subset of 124 drug compounds, comprises of both FDA approved drugs and under investigation chemical compounds, among the complete set of drugs in the GDSC data set is selected. These drugs are presented by their primary names that can be used to extract their SMILES by using Identifier Exchange Service of PubChem web interface (Kim et al., 2018). Then, these SMILES are used to obtain the fingerprints of the drugs and chemical compounds by the RCDK library in the R open-source software (i.e. R: A language and environment for Statistical Computing) (Guha et al., 2007; R Core Team, 2013). As a result, 10 different fingerprints are obtained. These fingerprints and their description are presented in the following. At first, each individual fingerprint was studied to obtain the features by HSIC-SCCA algorithm. However, the HSIC-SCCA algorithm was not successful to obtain features that maximize the correlation between obtained HSIC-SCCA variables. Therefore, we construct the drug compound data matrix by combining all 10 different fingerprints. After combining all 10 different fingerprints and removing duplicated features, the size of the variables is equal to 6103.

Standard fingerprint is 1024-bit fingerprint. This fingerprint is a path-based fingerprint. That is, each fingerprint is obtained by analyzing all different fragments of the molecule following a linear path. This path is limited to a certain number of bonds. Then every one of these paths is hashed to create the fingerprint. The length of the path is adjustable but the default length is 1024 (Cichońska et al., 2018; Guha et al., 2007; Cereto-Massagué et al., 2015).

Circular fingerprint is a 1024-bit fingerprint. This fingerprint, also called extended connectivity fingerprint (ECFP), is circular fingerprint with a maximum diameter equal to 6. In the circular type of fingerprints, instead of studying the chemical compounds in a linear path (starting from an atom and meet all other atoms of the chemical compound in a specific order), each arbitrary atom of a given chemical compound is considered as a center and the environment of this atom is studied up to a limited radius (Cichońska et al., 2018; Cereto-Massagué et al., 2015; Rogers and Hahn, 2010).

E-state fingerprint is a 79-bit fingerprint. Each bit corresponds to E-state, electro-topological state for atom electronic and topological characterization, properties of each atom in the molecular graph. E-state refers to the basic electronic state of an atom disturbed by the electronic influence of all other atoms in the molecule (Hall and Kier, 1995).

Extended fingerprint is a 1024-bit fingerprint. This fingerprint is a path-based fingerprint, similar to the standard fingerprint, but considers the connectivity between the atoms (Cichońska et al., 2018; Willighagen et al., 2017).

Graph fingerprint is a 1024-bit fingerprint. This fingerprint is path-based, similar to the standard fingerprint, but considers the connectivity between the atoms (Cichońska et al., 2018; Willighagen et al., 2017).

Hybridization fingerprint is 1024-bit fingerprint. This fingerprint is path-base, same as standard fingerprint, that only considers hybridization state (e.g. the angle between different atoms in the arbitrary chemical compound) and not aromaticity perception (Cichońska et al., 2018; Floris et al., 2014).

KR (KlekotaRoth) fingerprint is 4860-bit fingerprint. Each bit indicates the existence of a unique substructure. These unique substructures were generating by fragmenting a subset of Chambridge Diverse Set E library which had a positive effect on 24 cell-based examined assays (Klekota and Roth, 2008).

MACCS fingerprint is a 166-bit fingerprint. Each of these bits refers to a MACCS structural key that indicates different properties such as the number of atoms, bond, and custom properties. The mapping of these properties into a structure and consequently into the keybits is under software control (Keys, 2005; Durant et al., 2002).

PubChem fingerprint is an 881-bit fingerprint. Each bit refers to the presence of a substructure. These substructures set cover wild range of features, such as element counts, a different type of rings, atom pairs and many others (Bolton et al., 2008).

Short Path fingerprint is 1024-bit fingerprint. This fingerprint considers the shortest path between a pair of atoms. Additionally, it takes into account the ring systems and charges (Cichońska et al., 2018).

4.2 Pre-processing of Data sets

Due to the high number of variables in both drug compounds data set and cell lines data set, we perform a pre-processing step to reduce the number of features. This step was necessary in order to reduce the dimension of the data set and consequently reducing the consumed time for feature selection step.

In the cell line data set $\mathbf{X}_C$, the genes were selected by the variance of their expression values. At first, the variance of each gene expression overall cell line samples have been calculated. These values are in the range $[0, 18.92]$. The histogram of these values is illustrated in Figure 7.

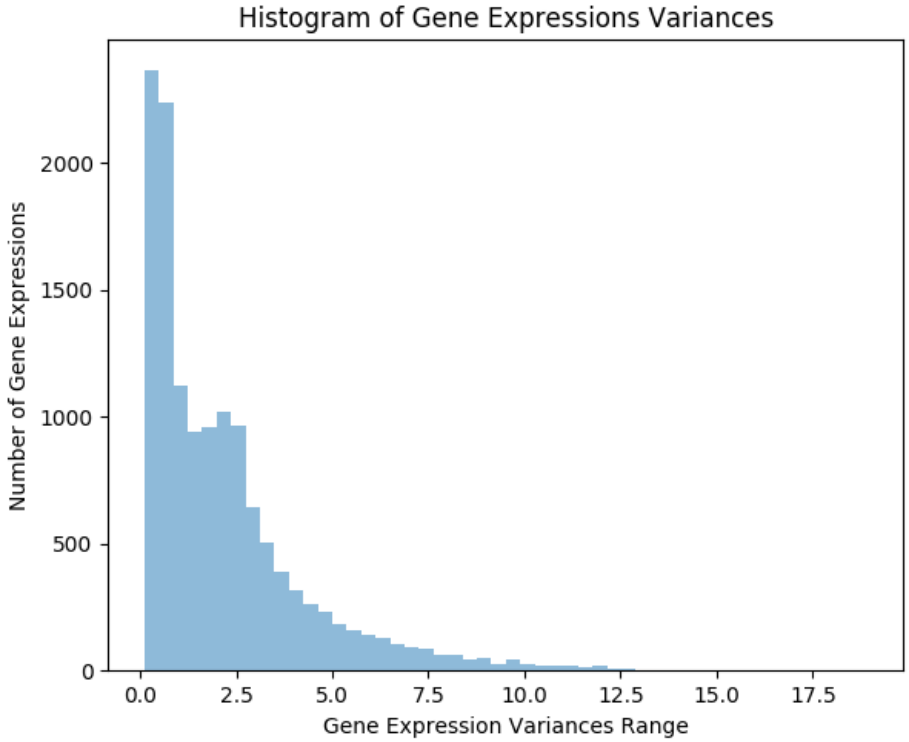


Figure 7: Histogram of gene expression variances.

As it can be seen in Figure 7, many of gene expression values have variances close to zero. Therefore, in the next step, the genes that have variances higher than 4 are selected. The number of genes that have variance higher than 4 is equal to 2067.

In the drug compound data matrix $\mathbf{X}_D$, the features are binary, 0, 1, values. Therefore, in order to remove the low variance features in this binary data set, the ratio of value one for each feature in all drug samples has been calculated. The histogram of the ratio of one in each feature is illustrated in Figure 8.

In Figure 8, on the x-axis, the ratio of the number of drugs that contain value one for an arbitrary variable to the total number of drugs is indicated. On the y-axis, the frequency of obtaining the same ratio for different variables are presented. Thus, on the x-axis values close to zero indicate that the frequency of values one in the corresponding

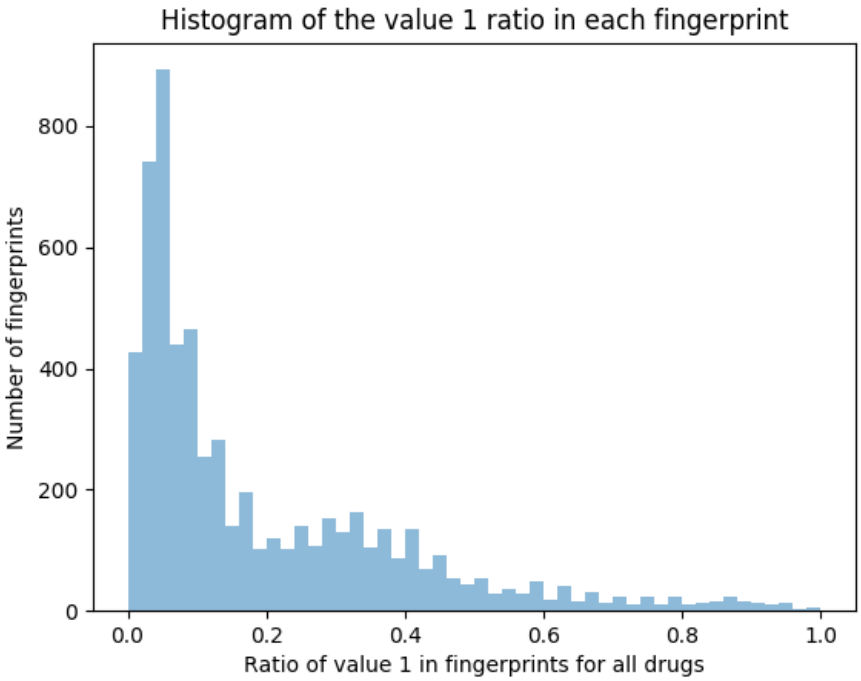


Figure 8: Histogram of the ratio of value 1 in each variable in drug data matrix.

variables are low and values close to one demonstrate that many of drug compounds have the value equal to one for the corresponding feature. Therefore, the variance among those features is low. The range $(0.4, 0.6)$ has been selected to improve the variance of features. The obtained data set comprises of 588 fingerprints as features.

4.3 Experiments

The drug sensitivity values are predicted by the pairwise kernel ridge regression method. In order to study the non-linear relations between samples and finding the meaningful model based on it, the Gaussian kernel has been used in both steps; feature selection, and regression prediction. This kernel is considered as one of the most popular candidates to study similarity-based prediction especially for the real values such as gene expressions of the cell lines. This kernel is presented in Equation 60.

$$k_x(\mathbf{x}^i,\mathbf{x}^j) = \exp(-\frac{\|\mathbf{x}^i-\mathbf{x}^j\|^2}{2\sigma^2}), \tag{60}$$

where $\mathbf{x}^i$ and $\mathbf{x}^j$ both are individual samples from cell line data set or drug compound data set. The parameter σ indicates the width of Gaussian kernel (Cichońska et al., 2018; Shawe-Taylor et al., 2004). The σ parameter is estimated by using Median heuristic. That is, the Euclidean distances between all pairs of samples are calculated and then the median of them has been selected as Gaussian width parameter (Flaxman et al., 2016).

4.3.1 Drug Sensitivity Prediction With Pairwise Kernel Ridge Regression

The pre-processed cell line and drug compound data matrices, $\mathbf{X}_C \in \mathbb{R}^{n\times p}$ and $\mathbf{X}_D \in \mathbb{R}^{n\times q}$ where $n = 124$, $p = 2067$ and $q = 588$, are used to predict the drug sensitivity values by adopting kernel pairwise ridge regression. The drug sensitivity data matrix, $\mathbf{Y}$, contains responses of 124 cell line samples to 124 drug compounds. Additionally, the Gaussian Kernel has been adopted to induce the non-linearity. That is, $\mathbf{K}_C$ and $\mathbf{K}_D$ indicate the cell lines and drug compounds Gaussian kernels and they are presented in the Equation 61.

$$\begin{aligned} k_C(\mathbf{x}_C^i,\mathbf{x}_C^j) &= \langle\phi(\mathbf{x}_C^i),\phi(\mathbf{x}_C^j)\rangle = \exp(-\frac{\|\mathbf{x}_C^i-\mathbf{x}_C^j\|^2}{2\sigma^2}) \\ k_D(\mathbf{x}_D^i,\mathbf{x}_D^j) &= \langle\phi(\mathbf{x}_D^i),\phi(\mathbf{x}_D^j)\rangle = \exp(-\frac{\|\mathbf{x}_D^i-\mathbf{x}_D^j\|^2}{2\sigma^2}), \end{aligned} \tag{61}$$

where $\mathbf{x}_C^i,\mathbf{x}_C^j$ refer to the individual cell lines, and $\mathbf{x}_D^i,\mathbf{x}_D^j$ indicate the individual drug compounds. ϕ refers to the mapping from input space to high-dimensional feature space for both cell lines and drug compounds data sets, $\phi : \mathbf{x}_D \in \mathscr{X}_D \rightarrow \phi(\mathbf{x}_D) \in \mathscr{H}_D$ and $\phi : \mathbf{x}_C \in \mathscr{X}_C \rightarrow \phi(\mathbf{x}_C) \in \mathscr{H}_C$. The σ value for both kernels are estimated by the Median heuristic. The Kronecker product of calculated drug kernel, $\mathbf{K}_D \in \mathbb{R}^{124\times 124}$, and cell line kernel, $\mathbf{K}_C \in \mathbb{R}^{124\times 124}$, is obtained for prediction by pairwise kernel ridge regression method. This kernel obtained by the Equation 62.

$$\mathbf{K} = \mathbf{K}_D \otimes \mathbf{K}_C = \begin{pmatrix} k_D(\mathbf{x}_D^1,\mathbf{x}_D^1)\mathbf{K}_C & k_D(\mathbf{x}_D^1,\mathbf{x}_D^2)\mathbf{K}_C & \cdots & k_D(\mathbf{x}_D^1,\mathbf{x}_D^n)\mathbf{K}_C \\ k_D(\mathbf{x}_D^2,\mathbf{x}_D^1)\mathbf{K}_C & k_D(\mathbf{x}_D^2,\mathbf{x}_D^2)\mathbf{K}_C & \cdots & k_D(\mathbf{x}_D^2,\mathbf{x}_D^n)\mathbf{K}_C \\ \vdots & \vdots & \ddots & \vdots \\ k_D(\mathbf{x}_D^n,\mathbf{x}_D^1)\mathbf{K}_C & k_D(\mathbf{x}_D^n,\mathbf{x}_D^2)\mathbf{K}_C & \cdots & k_D(\mathbf{x}_D^n,\mathbf{x}_D^n)\mathbf{K}_C \end{pmatrix}. \tag{62}$$

The kernel $\mathbf{K}$ can be used to obtain the solution for the pairwise kernel ridge regression when the $\mathbf{K}_D$ and $\mathbf{K}_C$ are the inner product between train samples of cell lines and drug compounds. The vector model parameter, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$, is calculated by the Equation 63.

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I}_N)^{-1} y, \tag{63}$$

where N is equal to the number of samples in cell lines train dataset, n_c, times number of samples in drug compounds train dataset, n_d, and $\mathbf{I}$ in $N \times N$ identity matrix, and λ is regularization parameter adjust the empirical error and model complexity ($\lambda > 0$). Then, the prediction for a test pair $(\mathbf{x}_D, \mathbf{x}_C)$ is calculated by the Equation 64.

$$f(\mathbf{x}_D, \mathbf{x}_C) = \sum_{l=1}^{N} \alpha^l k((\mathbf{x}_D^l, \mathbf{x}_C^l), (\mathbf{x}_D, \mathbf{x}_C)) = \alpha^T \mathbf{k}, \tag{64}$$

where $\mathbf{k}$ is a column vector with kernel values between train samples pair of cell lines and drug compounds, $(\mathbf{x}_D^l, \mathbf{x}_C^l)$, and test pairs, $(\mathbf{x}_D, \mathbf{x}_C)$.

In this study, we used the pairwise kernel ridge regression implemented in RLScore package. This implementation adopts several properties of the Kronecker product of two matrices for a fast and memory efficient calculation of pairwise kernel in the pairwise ridge regression. Therefore, this package enhances the model construction and prediction of large data sets (Pahikkala and Airola, 2016).

4.3.2 HSIC-SCCA As Feature Selection Step For Prediction Of Drug Sensitivity Values

The HSIC-SCCA method has been adapted, in order to obtain a sparse set of features for cell lines and drug compounds, that have a high correlation with drug sensitivity values in the non-linear space. In this experiment, each of cell line data set and drug compound data set, $\mathbf{X}_C$ and $\mathbf{X}_D$, has been paired with the drug sensitivity data matrix, $\mathbf{Y}_R$ separately and constructed the view A and view B of HSIC-SCCA method. Later, the projected vectors obtained by HSIC-SCCA method for cell line dataset and drug compound data set are used to calculate a sparse projection of features. The Gaussian kernel has been used to calculate the $\mathbf{K}_C$ and $\mathbf{K}_D$ from the sparse projection of features in the cell lines and drug compounds data set. At last, these kernels are used in the kernel pairwise ridge regression method to construct the predictive model. In the figure 9 a schematic presentation of the work-flow of the experiment is given.

The sparse projection vectors for drug compound data set, $\mathbf{u}_D \in \mathbb{R}^p$, $p = 588$, and cell line data set, $\mathbf{u}_C \in \mathbb{R}^q$, $q = 2067$, are obtained by solving the HSIC-SCCA optimization problem presented in the Equation 57 for both combinations of cell lines - drug sensitivity data matrices (e.g. $\mathbf{X}_C, \mathbf{Y}$), and drug compound - drug sensitivity data matrices (e.g. $\mathbf{X}_D, \mathbf{Y}$). These two optimization problems are presented in Equation 65.

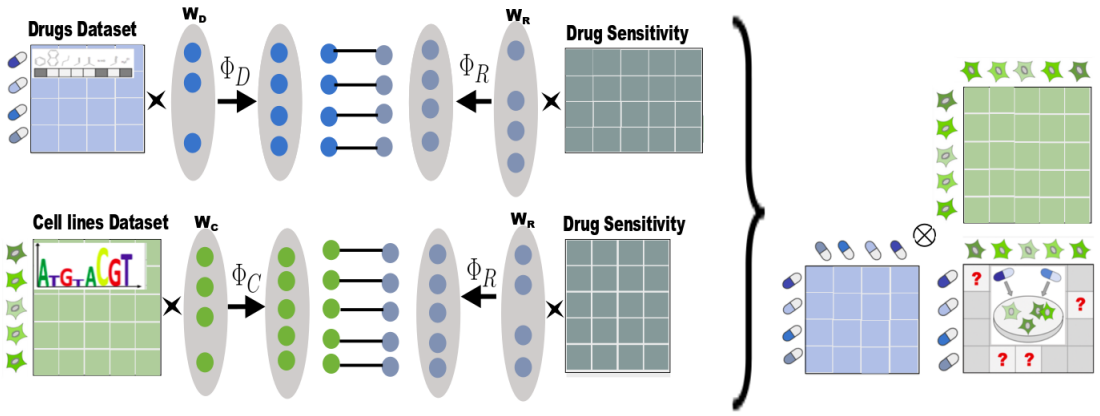


Figure 9: The work-flow for feature selection with HSIC-SCCA method and obtaining the kernel for selected features. $\mathbf{w}_R$, $\mathbf{w}_D$, and $\mathbf{w}_C$ are sparse projection vectors for drug sensitivity responses, drug compounds, and cell lines data matrices respectively. $\mathbf{\Phi}_R$, $\mathbf{\Phi}_D$, and $\mathbf{\Phi}_C$ are features maps for drug sensitivity responses, drug compounds, and cell lines data matrices respectively. These kernels are used in the construction of an optimal kernel pairwise ridge regression model to acquire the prediction for drug sensitivity values.

$$
\begin{aligned}
\max \quad & \rho(\mathbf{u}_C, \mathbf{v}_R) = \frac{trace(\hat{\mathbf{K}}_C^{u_C} \hat{\mathbf{K}}_R^{v_R})}{(n-1)^2}, \\
s.t. \quad & \|\mathbf{u}_C\|_1 \leq c_1 \text{ and } \|\mathbf{v}_R\|_1 \leq c_2 \\
\\
\max \quad & \rho(\mathbf{u}_D, \mathbf{v}_R) = \frac{trace(\hat{\mathbf{K}}_D^{u_D} \hat{\mathbf{K}}_R^{v_R})}{(n-1)^2}, \\
s.t. \quad & \|\mathbf{u}_D\|_1 \leq c_1^{'} \text{ and } \|\mathbf{v}_R\|_1 \leq c_2^{'}
\end{aligned}
\tag{65}
$$

where c_1, c_2 are optimized regularized parameter for the cell lines and drug sensitivity data matrices, and c_1', and c_2' are optimized regularization parameter for the pair drug compounds and drug sensitivity data matrices. These regularization parameter indicating the level of the sparsity of projection vectors for the pair of data sets under study. Gram matrices $\hat{\mathbf{K}}_C^{u_C}$, $\hat{\mathbf{K}}_C^{u_D}$, and $\hat{\mathbf{K}}_R^{v_R}$ indicate the centered Gram matrices that store non-linear similarity of projected cell lines, drug compounds, and drug sensitivity data sets. The Gaussian kernel has been used to obtain the non-linear relation between projected samples. The Equation 66 illustrate the exact formulation of these kernels.

$$
\begin{aligned}
k_C^{u_C}(\mathbf{x}_C^i,\mathbf{x}_C^j) &= \langle \phi(\mathbf{u}_C^{i^T}\mathbf{x}_C^i), \phi(\mathbf{u}_C^{j^T}\mathbf{x}_C^j)\rangle = \exp(-\frac{\|\mathbf{u}_C^{i^T}\mathbf{x}_C^i - \mathbf{u}_C^{j^T}\mathbf{x}_C^j\|^2}{2\sigma^2}),\\
k_D^{u_D}(\mathbf{x}_D^i,\mathbf{x}_D^j) &= \langle \phi(\mathbf{u}_D^{i^T}\mathbf{x}_D^i), \phi(\mathbf{u}_D^{j^T}\mathbf{x}_D^j)\rangle = \exp(-\frac{\|\mathbf{u}_D^{i^T}\mathbf{x}_D^i - \mathbf{u}_C^{j^T}\mathbf{x}_D^j\|^2}{2\sigma^2}),\\
k_R^{u_R}(\mathbf{x}_R^i,\mathbf{x}_R^j) &= \langle \phi(\mathbf{u}_R^{i^T}\mathbf{x}_R^i), \phi(\mathbf{u}_R^{j^T}\mathbf{x}_R^j)\rangle = \exp(-\frac{\|\mathbf{u}_R^{i^T}\mathbf{x}_R^i - \mathbf{u}_R^{j^T}\mathbf{x}_R^j\|^2}{2\sigma^2}).
\end{aligned}
\tag{66}
$$

After obtaining the Gaussian Kernels for projected drug compounds and cell lines data set, $\mathbf{K}_C^{u_C}$ and $\mathbf{K}_D^{u_D}$, these Kernel are used to obtain the pairwise kernel matrix $\mathbf{K}$ in the Equation 62. The optimized model for the kernels of projected data sets can be calculated by finding the vector of parameters $\boldsymbol{\alpha}$ given in the Equation 63 and the prediction is made by using this vector in the Equation 64.

4.3.3 KCCA As Feature Selection Step For The Prediction Of Drug Sensitivity Values

For selecting features in the cell line and drug compounds data sets with KCCA method, the canonical weight vectors are obtained in the high-dimensional Reproducing Kernel Hilbert Space (i.e feature spaces) $\mathscr{H}_C$ and $\mathscr{H}_D$. Unlike the HSIC-SCCA method, in the KCCA samples first, are mapped to high-dimensional feature space by using the kernels. Then, the canonical projection vectors are obtained by maximizing the correlation between canonical variables. The same as HSIC-SCCA step, the cell lines, and drug compounds data sets are pair individually with drug sensitivity data set and the canonical projection vectors of cell lines samples and drug compounds, $\boldsymbol{\alpha}_C$ and $\boldsymbol{\alpha}_D$, are obtained separately by solving two optimization problem given in the Equation 67.

$$
\begin{aligned}
&cos(\mathbf{z}_C,\mathbf{z}_R) = \max_{\mathbf{z}_C,\mathbf{z}_R\in\mathbb{R}^n} \langle \mathbf{z}_C,\mathbf{z}_R\rangle = \alpha_C^T\mathbf{K}_C^T\mathbf{K}_R\beta_R,\\
&\text{s.t} \quad \alpha_C^T(\mathbf{K}_C + c_1\mathbf{I})^2\alpha_C = 1 \ \text{ and } \ \beta_R^T(\mathbf{K}_R + c_2\mathbf{I})^2\beta = 1\\
&\\
&cos(\mathbf{z}_D,\mathbf{z}_R) = \max_{\mathbf{z}_D,\mathbf{z}_R\in\mathbb{R}^n} \langle \mathbf{z}_D,\mathbf{z}_R\rangle = \alpha_D^T\mathbf{K}_D^T\mathbf{K}_R\beta_R,\\
&\text{s.t} \quad \alpha_D^T(\mathbf{K}_D + c_1^{'}\mathbf{I})^2\alpha_D = 1 \ \text{ and } \ \beta_R^T(\mathbf{K}_R + c_2^{'}\mathbf{I})^2\beta^R = 1,
\end{aligned}
\tag{67}
$$

where $\mathbf{z}_C$, $\mathbf{z}_D$, $\mathbf{z}_R$ are canonical variables for cell lines, drug compounds and drug sensitivity data sets, respectively. $\mathbf{K}_C$, $\mathbf{K}_D$, and $\mathbf{K}_R$ refer to the kernel matrices sample in the cell lines, drug compounds, and drug sensitivity data sets. Additionally, c_1 and c_2 correspond to the regularization parameters for the pair cell lines and drug sensitivity input, and c_1' and c_2' indicate the regularization parameter for the pair drug compounds and drug sensitivity data matrices. The canonical vectors α_C, α_D and β_R are calculated by the solving Lagrange multiplier equations.

The solution of canonical vectors, for cell line data set α_C and for the drug compound data set α_D, that are calculated in the feature spaces of cell lines samples and drug compounds, are of the size $\mathbb{R}^n$ where $n = 124$. Therefore, in order to calculate the cell

lines and drugs compounds kernels of projected input data matrices, the obtained canonical variables vector $\mathbf{z}_C$ and $\mathbf{z}_D$ of all projected samples are used. The similarity between the samples in the high-dimensional feature space is obtained by Gaussian kernel and stored in the $\mathbf{K}_C$ and $\mathbf{K}_D$. The Equation 68 shows the calculation of these kernels

$$\begin{aligned} k_C^{\alpha_C}(\mathbf{x}_C^i, \mathbf{x}_C^j) &= \langle \phi(\mathbf{z}_C^i), \phi(\mathbf{z}_C^j) \rangle = \exp(-\frac{\|\mathbf{z}_C^i - \mathbf{z}_C^j\|^2}{2\sigma^2}) \\ k_D^{\alpha_D}(\mathbf{x}_D^i, \mathbf{x}_D^j) &= \langle \phi(\mathbf{z}_D^i), \phi(\mathbf{z}_D^j) \rangle = \exp(-\frac{\|\mathbf{z}_D^i - \mathbf{z}_D^j\|^2}{2\sigma^2}), \end{aligned} \tag{68}$$

where vector $\mathbf{z}_C^i$ and $\mathbf{z}_C^j$ store canonical variables of different components of individual samples in projected cell line data set, and $\mathbf{z}_D^i$ and $\mathbf{z}_D^j$ indicate the vector of canonical variables of different components in the projected drug compounds data set. Same as HSIC-SCCA, the pairwise kernel matrix $\mathbf{K}$ given as in the Equation 62 is obtain by Kronecker product of these Kernels and the optimal model is constructed by using this pairwise kernel in kernel ridge regression method.

4.3.4 General Protocols

In the all aforementioned experiments, the optimal model to predict the drug sensitivity values is obtained by the kernel ridge regression. The root mean-squared error (RMSE) has been used as model selection criteria for kernel ridge regression. The reason for such selection is that the RMSE follows the same scale as the original data set, thus it is often preferred over mean-squared error Hyndman and Koehler (2006). Additional to the RMSE criteria, Pearson correlation, Spearman's rank correlation, and c-index (concordance index) are reported for the generalization error of the optimal model.

Additionally, the regularization parameter of kernel ridge regression is selected by leave pair out 5×3 nested cross-validation technique. The term leaves pair out (LPO) refers to the fact that the cell line and drug compound pairs in the train data set are not shared with the test data set. Therefore, the pairs in the test dataset are not seen by the obtained optimal model on the train set. The LPO technique eliminates the bias factors in the model evaluation step Pahikkala et al. (2013). A grid search over 10^i for $i \in \{-3, -2, -1, 0, 1, 2, 3\}$ was performed to select the optimal regularization parameter in pairwise kernel ridge regression. Additionally, the same train and test samples are used in the outer folds of all experiments in order to make a more accurate comparison. In the next chapter, the results of these experiments are presented and discussed in details.

5 Results And Discussion

The objective of this thesis is to study the effect of feature selection in drug sensitivity. In order to accumulate the features, which increase the correlation of drug compounds and cell lines with drug sensitivity, two methods HSIC-SCCA and KCCA, has been adopted. The accuracy of the obtained features is evaluated by the accuracy of the prediction made by the optimal model. In order to construct the optimal model, the pairwise kernel ridge regression has been adopted. In the following, the results of three different experiments, pairwise kernel ridge regression, HSIC-SCCA as feature selection step, and KCCA as feature selection step on the prediction of the drug sensitivity is presented. First, we present the results of the prediction of the optimal model obtained by pairwise kernel ridge regression method on the re-processed data. Then the effect of feature selection by HSIC-SCCA, for pre-processed data, is demonstrated. In the end, the KCCA has been employed as an alternative feature selection technique to compare with the HSIC-SCCA method.

5.1 Drug Sensitivity Prediction With Pairwise Kernel Ridge Regression

In this experiment, the pairwise kernel is constructed by the Kronecker product of two Gaussian kernels. The cell line kernel $\mathbf{K}_C \in \mathscr{H}_C$ and drug kernel $\mathbf{K}_D \in \mathscr{H}_D$. This pairwise kernel is used for obtaining the optimal model by pairwise kernel ridge regression method. A nested 5×3 cross-validation partitions the input data sets into the 5 different outer folds randomly. Therefore, 4 train folds comprise of 100 cell lines and drug compounds samples, and one of the train folds contains 99 samples. In each fold, the optimal regularization parameter has been selected based on averaged accuracy of 3 inner folds. Then, the obtained regularization parameter has been used for training of the model for all the train samples in the outer fold. At last, this optimal model has been used for predicting the drug sensitivities in the outer test fold.

There is a great difference between the distributions of the train and test folds in each iteration. This difference is perceptible in Figure 10. The top row of the figure illustrates the distribution of the drug sensitivity values in the different train and test folds in two histograms. The top left histogram presents the labels in the train folds and the top right plot shows the distributions of label values for test folds. The different folds are indicated by different colors. The two scatter plots on the bottom row indicate the correlation between the prediction of the optimal model, on the x-axis, and the real values, on the y-axis, of the train (bottom left) and test (bottom right) folds. As it can be seen in the bottom left scatter plot, the model is able to learn from the train samples but this obtained model has not achieved high performance on the test samples. This can be explained by the wild differences between the train folds and test folds. Thus, due to these differences, the optimal models learned from the train folds cannot achieve high performance on the test folds.

As It is presented in Figure 10, the average RMSE obtained for the model obtained by pairwise kernel ridge regression method is equal to 2.96 with standard deviation equal to 0.17. The overall C-index of the model is equal to 0.62 with standard deviation

equal to 0.05. The Pearson correlation value for this model is equal to 0.45 with standard deviation equal to 0.14 and the Spearman correlation criteria is equal to 0.35 with standard deviation equal to 0.15. The high standard deviation over 5 folds for all accuracy criteria is strong evidence of the high diversity between samples of the train and test folds. In the next section, we want to reduce this diversity by adopting HSIC-SCCA technique as a feature selection step.

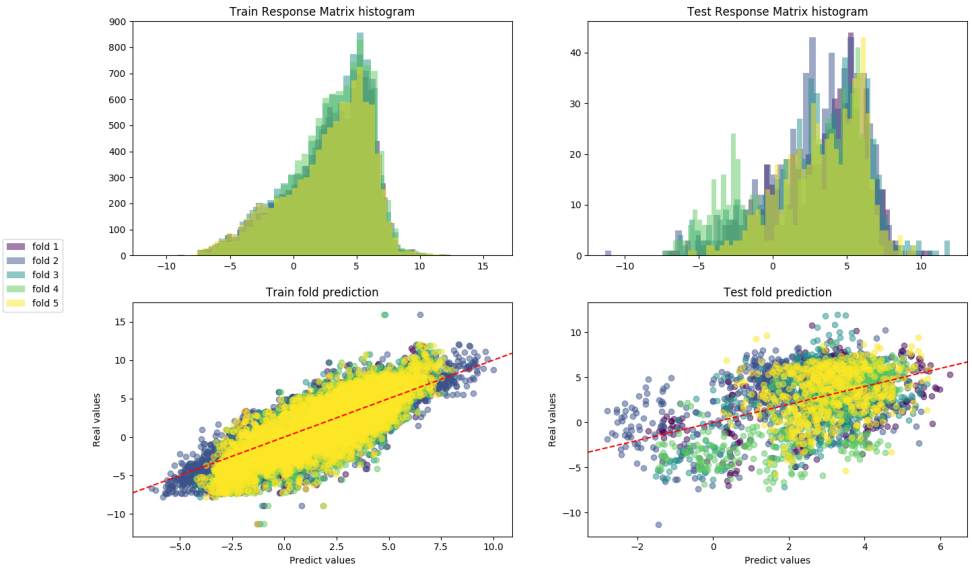


Figure 10: Pairwise ridge regression prediction for train and test folds of the pre-processed cell line and drug compounds data-sets. Each fold is indicated by a different color. The top left plot illustrates the distribution of the drug sensitivity values for train folds, the top right plots indicate the distribution of the same measurements in the test folds. The two scatter plots in the bottom row illustrating the prediction of the optimal model for train folds (left) and test folds (right).

5.2 The effect of HSIC-SCCA as Feature Selection Step

In this part, we explain the result of HSIC-SCCA method as a feature selection step. We study the effect of selected features from cell lines and drug compounds data matrices that have a high correlation with drug sensitivity values in non-linear feature space. These features are selected by a sparse canonical vector and stored in the canonical variables. The HSIC-SCCA method benefits from stochastic projected gradient descent algorithm and finds the optimal projection vector by selecting small batches of samples in iterative steps. This algorithm initiates with a projection vector randomly. Through our experiments, we noticed that the random initialization of the projection vector has a dramatic effect on the convergence of the optimizer. In other words, this algorithm can converge to the different local minimal depending on the random selection of

the initial projection vector. In order to overcome this difficulty, the number of random initialization of the projection vectors was set to 30. Other criteria for obtaining optimal projection vector were the maximum number of iterations and convergence limit. Selecting small values for each of these parameters can cause the algorithm to suffer from an early stopping point. Therefore, I set the convergence limit, ϵ to 10^{-7}, and the maximum number of iteration to 500. The learning curve of best projection vector obtained by the HSIC-SCCA method in each fold for both pairs of cell lines - drug sensitivity data matrices ($\mathbf{X}_C$, $\mathbf{Y}_R$), and drug compounds - drug sensitivity data matrices ($\mathbf{X}_D$, $\mathbf{Y}_R$) are presented by the mean and standard deviation over 5 train folds in the Figure 11. All the learning curves in Figure 11 illustrate the first component obtained by HSIC-SCCA for drug compounds data matrix (left) and cell lines data matrix (right).

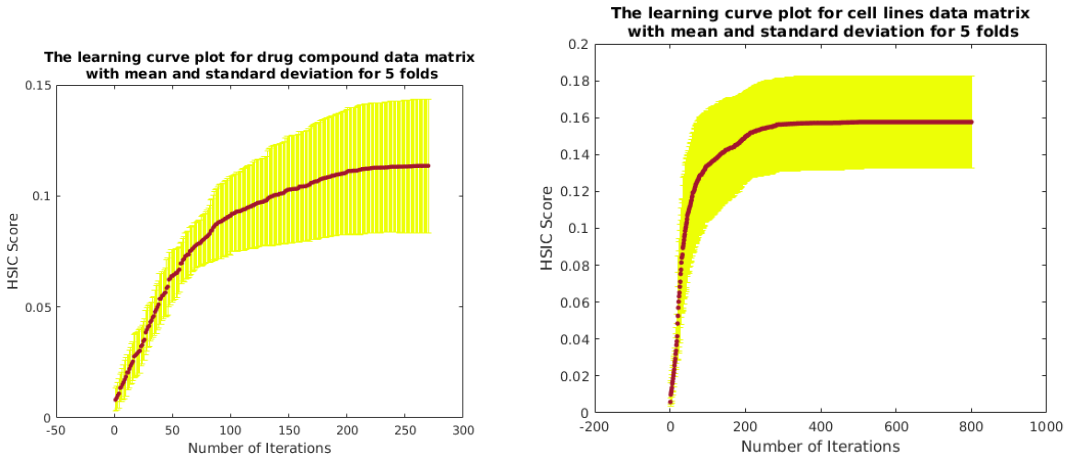


Figure 11: The overall learning curve of the best HSIC-SCCA projection vectors achieved for the first components in different 5 train folds by the mean and standard deviation over 5 train folds. The left plot presents the learning curves for the first components of five different train folds for drug compounds data matrix. The right plot illustrates the learning curves for the first components of five different train folds for cell lines data matrix.

Figure 11 illustrates that in most of the cases, the algorithm stops before the maximum iteration has been met. That is due to the fact that the minimum convergence criteria have been met. As it can be seen in Figure 11 in the different folds, the Algorithm has converged to the different local minima. This is because of two reasons. First, the diversity of samples between different folds that causes the selection of different optimal projection vectors. Second, the randomly initializing of projection vectors. The effect of these two factors can be observed more vividly in Figure 12. In this figure, the HSIC-SCCA score for train and test folds of 10 different components over five folds is presented by their mean and standard deviation.

In Figure 12, the left plot illustrates the results for drug compounds data matrix and the right plot presents the results of cell lines data matrix. Ideally, the HSIC-SCCA score of different components should reduce as we are removing stronger variables in each obtained component by deflation techniques. However, the plots in Figure 12 illustrate

different reality. This is due to the fact that random initialization point causes the algorithm to fall in different local minima. Therefore, the subset of features selected in different components is equally good in most of the time. The higher range of standard deviation is another factor that demonstrates the effect of random initialization for the projection vector and diversity among the samples in different folds, as high standard deviation indicates the high difference between HSIC-SCCA value for a component among different folds.

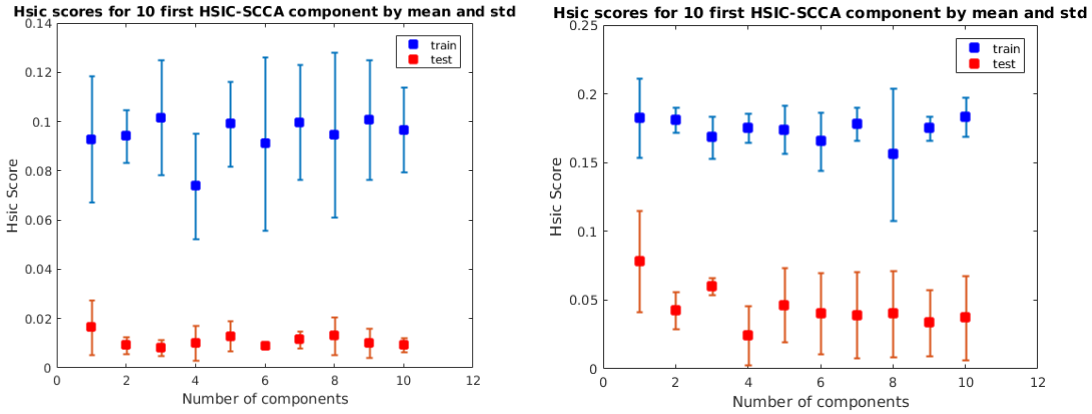


Figure 12: The illustration of HSIC-SCCA scores for 10 different components over 5 outer folds of cross-validation by their means and standard deviations.

The HSIC-SCCA variable for each component is obtained by inner product between the data matrices and the optimal projection vectors in the different folds. In order to study the accuracy and effect of these projection vectors on the different train and test folds, we plot each of these HSIC-SCCA variables in a score plot. The score plot is a 2D dimensional presentation of canonical variables, where one axis refers to the obtained canonical variables of one view (such as drug compounds or cell lines data matrices), whereas the other axis presents the canonical variables of the other view (drug sensitivity data matrix). These results are presented in Figure 13. In this figure, the first component of different folds is presented in different colors. The first row of this plot presents the results of canonical variables for the train (left) and test (right) folds of the pair drug compound data matrix (x-axis) and drug sensitivity data matrix (y-axis). The second row of this plot illustrates the canonical variables of cell lines data matrix (x-axis) and the drug sensitivity data matrix (y-axis).

The obtained canonical variables yield high correlation in the non-linear feature spaces, therefore, the aim of this plot is not illustrating the correlation between two canonical variables of each view of a plot, but to studying whether the projection vector successfully captures the similar pattern in the train and test folds. As it can be seen in Figure 13, the similar pattern between the same train and test of each fold is obtained. That is, the HSIC-SCCA was able to capture the strong variables in each fold that was equally explanatory in the test folds. The HSIC scores for the cell lines data set are higher than the HSIC score of drug compounds data set in general. This is due to the higher number of features that have a stronger correlation with drug sensitivity.

The differences between the HSIC scores of train and test folds is due to the high difference between samples of the train and test folds. Even though the HSIC-SCCA algorithm managed to extract the features that result in a high correlation between two views of a training fold, but some of these features could have really low values in the understudying test fold. Therefore, the test values of the HSIC score are lower in test folds.

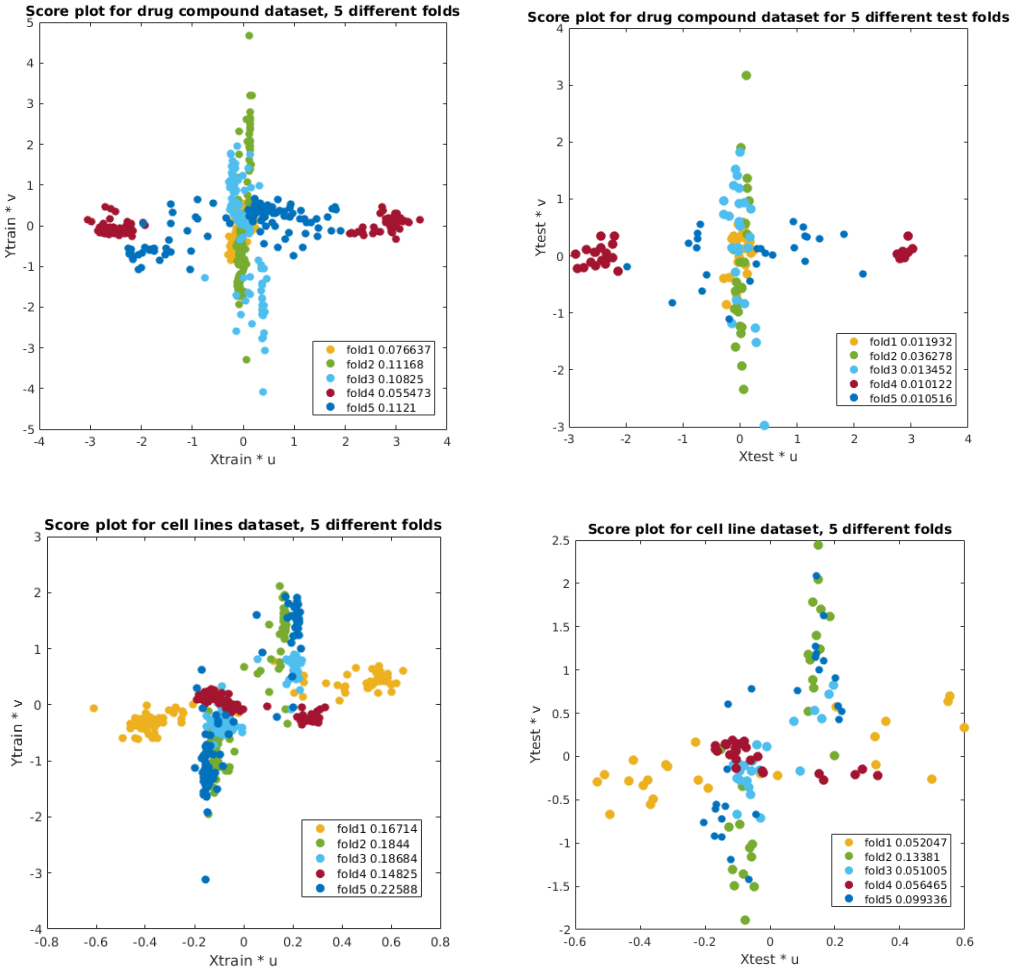


Figure 13: Score plots for pairs drug compound data matrix and drug sensitivity (first row), the x-axis refers to the drug compound variables and the y-axis refers to the drug sensitivity variables. The pairs consist of cell lines data matrix and drug sensitivity are presented in the (second) row, the x-axis refers to the cell line variable and the y-axis refers to the drug sensitivity variables. The left plots in both rows refer to the training folds and the right plots refer to the test folds.

Now we want to see the effect of the feature selection with HSIC-SCCA algorithm on the prediction of drug sensitivity values. The same as the previous experiment, the pairwise kernel ridge regression in the RLScore package has been adopted for prediction of drug sensitivity values. The HSIC-SCCA variables of cell lines data matrix and drug

compound data matrix obtained by the first component has been chosen and the cell lines kernel, $\mathbf{K}_C$, and drug compounds kernel, $\mathbf{K}_D$ are obtained by applying the Gaussian kernel on these HSIC-SCCA variables. These kernels were used by the pairwise kernel ridge regression method to obtain the optimal model. The prediction results of pairwise kernel ridge regression are illustrated in Figure 14.

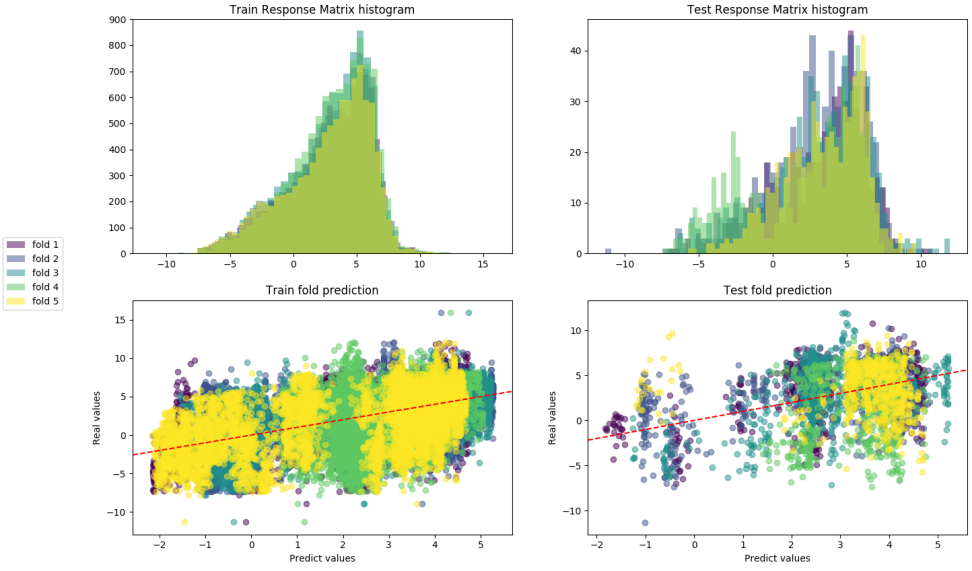


Figure 14: Pairwise ridge regression prediction for train and test folds of cell lines and drug HSIC-SCCA canonical variables. Each fold is indicated by a different color. The top left plot illustrates the distribution of the drug sensitivity values for train folds, the top right plots indicate the distribution of the same measurements in the test folds. The two scatter plots in the bottom row illustrating the prediction of the optimal model for train folds (left) and test folds (right).

As it can be observed in Figure 14, the accuracy of the prediction obtained by HSIC-SCCA variables is lower than the original data matrices. The RMSE of the prediction increased to 3.13 with higher standard deviation equal to 0.43, the C-index value dropped to 0.56 with standard deviation equal to 0.62, the both Pearson correlation and Spearman rank correlation have decreased to 0.38 with standard deviation 0.28 and 0.18 with standard deviation equal to 0.18 respectively. This is due to the fact that differences between the train and test folds have increased by adopting HSIC-SCCA. As we pointed it out previously, due to the fact of random initialization of projection vector and high diversity between samples, the HSIC-SCCA algorithm has converged to different local minima, therefore, different variables have been chosen by the algorithm among all features. That is, the sparse subset of features obtained by the algorithm in the different training folds cannot extract the strong features in the test fold. Although the score plots in the Figure 13 illustrate the similar pattern between the train and test folds, still the HSIC-SCCA canonical vectors increase the diversity between the samples, therefore; the

accuracy of the prediction of pairwise kernel ridge regression is reduced.

In order to reduce the diversity between the samples of different folds after the feature selection, a rotational technique has been adopted. In this technique, the HSIC-SCCA variable of one fold has been chosen as a destination and all other HSIC-SCCA variables in the other folds have been rotated in order to reduce the distances between HSIC-SCCA variables of all folds. The rotation matrix for different folds is calculated by Equation 69.

$$\mathbf{Q} = \mathbf{U}^{(1)}\mathbf{X}^T\mathbf{X}\mathbf{U}^{(2)T}(\mathbf{U}^{(2)}\mathbf{X}^T\mathbf{X}\mathbf{U}^{(2)T})^{-1}, \tag{69}$$

where $\mathbf{Q}$ refers to the rotation matrix for each fold, the normalized destination projection matrix is indicated by $\mathbf{U}^{(1)} \in \mathbb{R}^{p\times t}$ where p is equal to the number of features in the corresponding data matrix and t is the number of components. $\mathbf{X}$ is the given data matrix and $\mathbf{U}^{(2)}$ present the normalized projection data matrix that we want to rotate. This rotation is applied to the projection vectors obtained in different folds for both views of pair cell lines - drug sensitivity and drug compounds - drug sensitivity by the inner product of rotation matrix $\mathbf{Q}$ and the projection matrices $\mathbf{U}^{(2)}$. By adopting this technique we achieved more similar HSIC-SCCA variables for different folds. This rotation has been applied to HSIC-SCCA variables of the drug sensitivity data matrix in order to be able to illustrate the result in a score plot. These score plots are presented in Figure 15.

In Figure 15, all the projection vectors of pair cell lines - drug sensitivity data matrices are rotated to the first fold, and for the pair drug compound - drug sensitivity data matrices all projection vectors are rotated to the fifth fold. This is because of the fact that these folds illustrate a higher correlation between two views in the score plot of Figure 13. Additionally, the better rotation was obtained when 3 first HSIC-SCCA components were considered for all folds. However, the rotation results yield better score plots for the pair cell lines - drug sensitivity data matrices, the second row in Figure 15, for both train and test folds. on the other hand, the rotation of the canonical variables of different folds for the pair drug compounds - drug sensitivity data matrices scattered the HSIC-SCCA canonical variables, especially on the test folds. This can be due to the fact that the destination HSIC-SCCA variable, the canonical variable of the fold five that we want to rotate all HSIC-SCCA variables of other folds to it, have more scattered scatter plot in the Figure 13, even though it has high HSIC score.

In Figure 16 the accuracy of the prediction of the optimal model based on the rotated HSIC-SCCA variables, obtaiened on the cell line and drug compound data matrices, is presented. As can be seen in this figure, the accuracy of the prediction for the test samples has slightly improved. The overall RMSE reduced to 2.83 with standard deviation equal to 0.24, C-index slightly improved and reached to 0.63 with less standard deviation 0.4, Pearson correlation increased to 0.5 with lower standard deviation equal to 0.1, and the Spearman correlation reached to 0.39 with standard deviation equal to 0.1. The improvement of the results is due to the fact that the rotation of the projection vectors decreases the diversity among the samples in the different folds, therefore, the optimal model obtained from the training fold has a higher accuracy of the test fold.

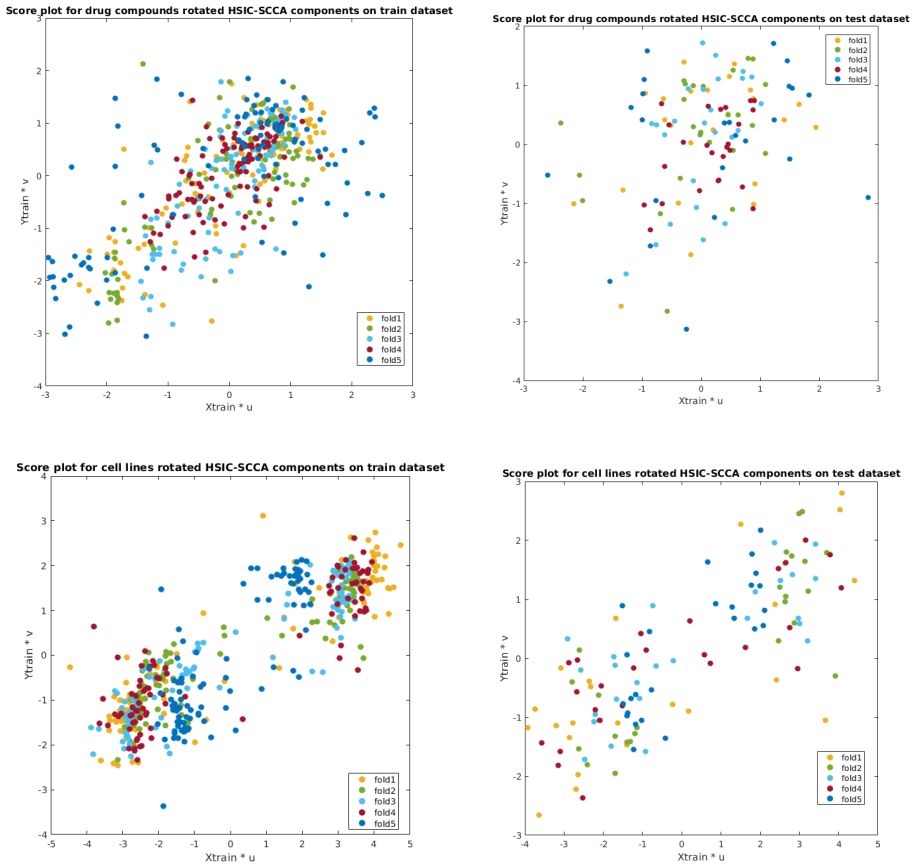


Figure 15: Score plots of pairs drug compounds - drug sensitivity data matrices (first row) and the cell lines data - drug sensitivity data matrices (second row) after rotation of projection vectors. The left plots in both rows refer to the training folds and the right plots refer to the test folds.

These results indicate that the low prediction accuracy for features selection by HSIC-SCCA is due to the High diversity among samples of different folds after projection. The first row of Figure 16 presents the histogram of the drug sensitivity values in different train (left) and test (right) fold. In the second row of this figure, the correlation between the prediction values (x-axis) and real values (y-axis) of the different train (left) and test (right) folds are illustrated.

5.3 The Effect of KCCA as Feature Selection Step

Based on the observation made on the feature selection with HSIC-SCCA, we decided to evaluate the effect of non-sparse feature selection method on the prediction of drug sensitivity values. The KCCA has been selected as our comparison model. In this model, a regularised parameter is optimized on the train samples to prevent the model from obtaining the canonical correlation equal to one. In order to obtain these parameters, a 5-fold cross-validation technique has been used. After obtaining the regularization

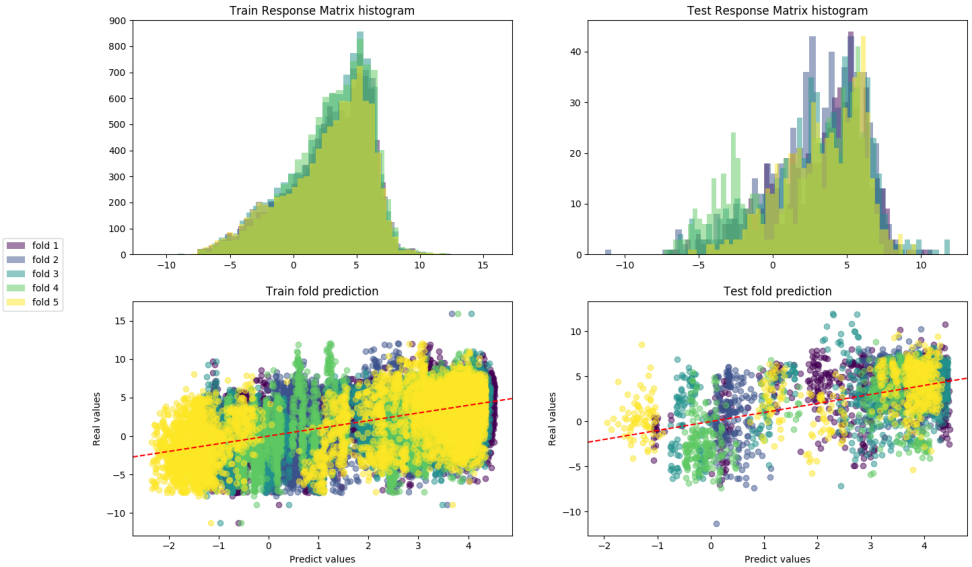


Figure 16: Pairwise ridge regression prediction for train and test folds of the pre-processed cell line and drug compounds data-sets. Each fold is indicated in a different color. The top left plot illustrates the distribution of the drug sensitivity values for train folds, the top right plots indicate the distribution of the same measurements in the test folds. The two scatter plots in the bottom row illustrate the prediction of the optimal model for train folds (left) and test folds (right).

.

parameter, these values are used to obtain an optimal model. Then this model is applied for the complete data sets to obtain the canonical variable. The experiment workflow is the same as the implementation of KCCA in the paper “A Tutorial on Canonical Correlation " (Uurtio et al., 2018b).

In the feature selection step for both data matrices in the pair drug compounds - drug sensitivity data matrices a grid search over the value of the range $\{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$ has been made. The selected value for the drug compound data set is equal to 3 and this value for the drug sensitivity data matrix is equal to 2.5. The canonical correlation obtains for these regularized parameters on the train data set is equal to 0.2014, which indicates that a high correlation between two data matrices of this pair could not be found. These regularized parameters are adapted to construct a model and this model has been applied to the complete data matrices in both views. The canonical correlation of the first component for the pair drug compound - drug sensitivity data matrices is equal to 0.63.

parameter for the data matrices of the pair cell lines- drug sensitivity data matrices. The regularized parameter chosen by the algorithm for this pair data matrices is equal to 0.5 and 0.5, for cell line data matrix and drug sensitivity data matrix respectively. Since the smallest values in the range have been chosen by the algorithm for both views, a range constructed of the values in the set 10^i where $i \in \{-3, -2, -1, 0\}$ has been tested.

The same as the previous experiment, the smallest value, 10^{-3}, has been chosen by the algorithm. The canonical correlation for the train folds for both experiments is equal to one. That is, due to a large number of features in the cell line data matrix, the obtained canonical correlation scores for all the combinations of the values in the grid search range are equal to one. This is due to the fact that the system under study is an undetermined system. Additionally, by choosing the smallest tested values, 0.001, for both regularized parameter the canonical correlation of the first component obtained for the pair cell line - drug sensitivity data matrices is equal to one. On the other hand, this value for the regularized parameters set to 0.5 in the same pair is equal to 0.95. This indicates that the regularized parameter 0.5 can overcome the undetermined system better; therefore, the value 0.5 has been chosen for the regularized parameter. The score plot of the first component obtain by this procedure is illustrated in Figure 17.

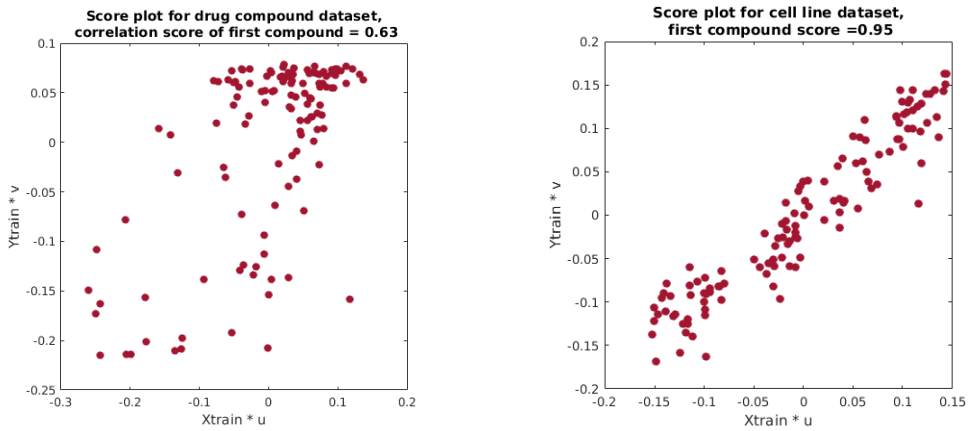


Figure 17: Score plots of two pairs drug compounds - drug sensitivity data matrices (left) and the cell line - drug sensitivity data matrices (right) for the first component obtained for each pair of data matrices by KCCA algorithm.

The left score plot in Figure 17 present the pair drug compounds - drug sensitivity data matrices. The pair of cell lines - drug sensitivity data matrices are presented in the right score plot. On the other hand, the KCCA method has not achieved a high correlation between the drug compounds and drug sensitivity data matrices. This can be due to the high diversity between features of different samples. The correlation score of 5 different KCCA components has been obtained and presented in Figure 18.

As it can be seen in Figure 18, the correlation score between two successive components of the same pair is decreasing. This is due to the fact that these components are obtained by the eigenvalue decomposition of two kernel matrices obtained from the Equation 67; therefore, they are in descending order. The accuracy of the prediction for the 5 variables obtained by adopting KCCA for feature selection is presented in Figure 19.

As it can be seen in Figure 19, adopting KCCA for the feature selection yield a similar result as original data in the regression prediction step. The RMSE obtained by this feature selection method is equal to 2.96 with the standard deviation equal to ± 0.17, the

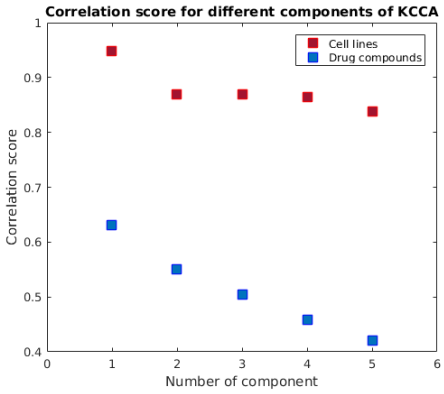


Figure 18: Correlation scores for 5 first components of KCCA obtained from drug compound - drug sensitivity data matrices (blue), and cell line - drug sensitivity data matrices (red).

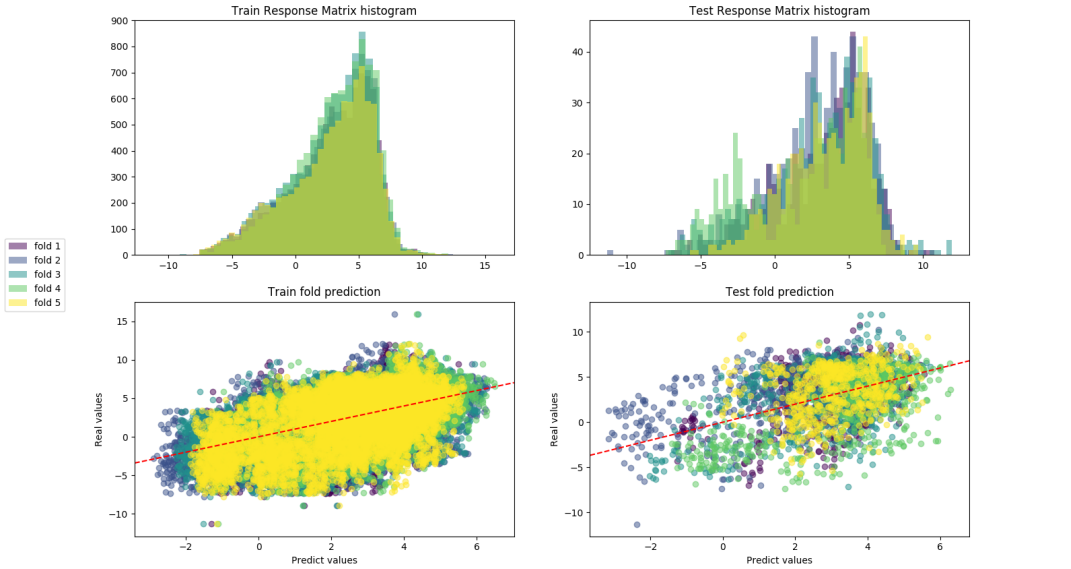


Figure 19: The accuracy of prediction obtained by five first components of KCCA obtained from the pair drug compound - drug sensitivity data matrices, and the pair cell line - drug sensitivity data matrices. The first row presents the histogram of the drug sensitivity values for the different train (left) and test (right) folds. The second row of the plot illustrates the correlation between the prediction values (x-axis) and real value (y-axis) on the train folds (left) and the test folds (right).

C-index is equal to 0.63 with standard deviation equal to ±0.04. The obtained Pearson correlation is equal to 0.46 with standard deviation equal to ±0.13 and the Spearman

correlation is equal to 0.39 with standard deviation equal to ±0.1. Although these results are similar to the prediction made with pairwise kernel ridge regression without feature selection step, they are better than the result obtained by adapting HSIC-SCCA method for feature selection. This is because of the fact that the samples of the different folds have lower variances compare to the case of HSIC-SCCA, after feature selection step. The similarity between the result of this experiment with the experiment where no feature selection step has been considered is because of the fact that by considering five different canonical variables obtained from the KCCA for constructing the cell line and drug compound kernel most of the features in both data set has been considered. Therefore, the regression prediction for drug sensitivity values is similar for these two experiments. In order to eliminate this problem, a smaller number of canonical variables for both cell line and drug compounds data matrices has been selected.

By choosing the first three KCCA canonical variables for both drug compounds and cell lines data matrices, the similar prediction accuracy, $RMSE = 2.96 \pm 0.16$, as before has been obtained. However, when the number of chosen variables decreases to one, the result of prediction has improved slightly. The RMSE obtained by first KCCA variables is equal to 2.90 with the standard deviation equal to ±0.19, the C-index is equal to 0.63 with standard deviation equal to ±0.03. The Pearson correlation and Spearman correlation are equal to 0.48 ± 0.1 and 0.39 ± 0.09 respectively. That is, the prediction obtained by the first KCCA canonical variable result in better prediction than the model build on by more than one canonical variables. This prediction result is better than the result without feature selection, therefore; we can conclude that the first canonical variables of both drug compound and cell line data matrices are more correlated with the drug sensitivity values. Additionally, the result of first KCCA canonical variable is not higher than the result obtained by using rotated HSIC-SCCA canonical variables. This is due to the fact that in the KCCA method, use l_2-norm regularized constraints, therefore; all the features are selected in the canonical variables but they have different weights that correspond to their ability to increase the correlation between two views in the non-linear space. On the other hand, in the HSIC-SCCA method, a sparse set of features is selected by l_1-norm regularized constraints. That is, this method is able to eliminate noise in the data matrices by choosing those features that improve the correlation between two views in the non-linear space. The summary of the results of all experiments is presented in Table 1.

Experiment	RMSE	C-index	Pearson correlation	Spearman correlation
Pairwise KRR	2.96 ± 0.17	0.62 ± 0.06	0.45 ± 0.14	0.35 ± 0.15
HSIC-SCCA	3.13 ± 0.43	0.56 ± 0.06	0.31 ± 0.21	0.18 ± 0.18
Rotated HSIC-SCCA	2.83 ± 0.24	0.62 ± 0.04	0.50 ± 0.10	0.39 ± 0.10
KCCA (5 components)	2.96 ± 0.17	0.63 ± 0.04	0.46 ± 0.13	0.39 ± 0.10
KCCA ($1^{st} component$)	2.90 ± 0.19	0.63 ± 0.03	0.48 ± 0.10	0.39 ± 0.09

Table 1: Table of results for five experiments. The pairwise KRR refers to the result obtained by using pairwise kernel ridge regression for cell line and drug compound data matrices. The HSIC-SCCA indicates the result of selecting cell line and drug compounds features by using HSIC-SCCA method. The rotated HSIC-SCCA refers to the results obtained by rotated canonical variables that are obtained by HSIC-SCCA from cell line and drug compound data matrices. The KCCA (5 components) refers to the results obtained by using five first components obtained KCCA method for cell line and drug compound data matrices. The KCCA ($1^{st} components$) indicates the results obtained by the first canonical variable obtained by KCCA from cell line and drug compounds data matrices. All the results are presented by their means and standard deviations over five outer folds.

6 Conclusion

The objective of this thesis was to evaluate the performance of non-linear variation of canonical correlation analysis, specifically Kernel CCA and HSIC-SCCA, in the feature selection for prediction of the drug sensitivity values. Previous works indicate that kernel methods are able to provide an accurate prediction by extracting non-linear relation among features of input data sets. Additionally, this performance improves when both cell lines and drug compounds data matrices are considered for obtaining the model. Therefore, we applied a non-linear variation of the canonical correlation analysis to extract features from both cell lines and drug compounds data matrices that are highly correlated with drug sensitivity values in non-linear space. The aim of this step was to improve the prediction accuracy by removing irrelevant features. Later, these features are used to obtain a model by using pairwise kernel ridge regression method.

As it has been indicated in the result section, when working with the underdetermined system, where the number of features is extremely higher than instances, the performance of the predictive model can be affected greatly. This effect can be more clear when the variance among the samples is high. In our study, the number of samples for both cell lines and drug compounds data matrices are equal to 124, whereas, the number of features in the pre-processed cell lines data set is equal to 2067 and the number of feature in the pre-processed drug compounds data matrix is equal to 588. Additionally, we observe a great difference between the distributions of drug sensitivity values for different 5 folds.

All these facts, affect the performance of the predictive model. The benchmark model, pairwise kernel ridge regression on the pre-processed data, achieved the accuracy of RMSE equal to 2.96, and C-index equal to 0.62. However, the feature selection with HSIC-SCCA was not able to improve this result. The obtained accuracy by this model is equal to $RMSE = 3.13$ and $C - index = 0.56$. This can be explained by the fact that the differences between samples of different folds have increased by adopting HSIC-SCCA method as feature selection. In order to improve the result, we adopted a rotation matrix obtained by three first canonical vectors of HSIC-SCCA to rotated the canonical variables of different folds to the same origins. This technique improved the accuracy of the prediction, and the RMSE of the obtained model reduced to 2.83 and C-index criteria improved to 0.62.

As a comparison, we study the effect of KCCA as feature selection step. The results of the predictive model obtained from five first KCCA variables was the same as the model obtained on the original input data matrices. This is because of the fact that considering five first KCCA variables combines all of the features, therefore; no improvement has been observed. Hence, using only the first variable obtained by KCCA for both cell lines and drug compounds slightly improved the accuracy of the obtained model. The RMSE of this model is equal to 2.90 and the C-index is equal to 0.63.

Finally, the results suggest that sparse set of features obtained in the non-linear feature space, by adapting HSIC-SCCA, can improve the prediction of the drug sensitivity values but in some cases such as this study they can increase the difference among the instances of different folds obtained by cross-validation technique. Thus, the accuracy of the predictive model is reduced. Additionally, HSIC-SCCA can perform better in the

cases that difference between the number of instances and the number of variables is not high.

References

MA Aizerman. The probability problem of pattern recognition learning and the method of potential functions. *Automation and Remote Control*, 25:1175–1193, 1964.

S AKAHO. A kernel method for canonical correlation analysis. In *Proceedings of the International Meeting of the Psychometric Society (IMPS2001)*. Springer-Verlag, 2001.

Muhammad Ammad-Ud-Din, Elisabeth Georgii, Mehmet Gonen, Tuomo Laitinen, Olli Kallioniemi, Krister Wennerberg, Antti Poso, and Samuel Kaski. Integrative and personalized qsar analysis in cancer by kernelized bayesian matrix factorization. *Journal of chemical information and modeling*, 54(8):2347–2359, 2014.

Muhammad Ammad-ud din, Suleiman A Khan, Disha Malani, Astrid Murumägi, Olli Kallioniemi, Tero Aittokallio, and Samuel Kaski. Drug response prediction by inferring pathway-response associations with kernelized bayesian matrix factorization. *Bioinformatics*, 32(17):i455–i463, 2016.

Theodore Wilbur Anderson, Theodore Wilbur Anderson, Theodore Wilbur Anderson, Theodore Wilbur Anderson, and Etats-Unis Mathématicien. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255, 2013.

Wael Awada, Taghi M Khoshgoftaar, David Dittman, Randall Wald, and Amri Napolitano. A review of the stability of feature selection techniques for bioinformatics data. In *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*, pages 356–363. IEEE, 2012.

Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603, 2012.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Evan E Bolton, Yanli Wang, Paul A Thiessen, and Stephen H Bryant. Pubchem: integrated platform of small molecules and biological activities. In *Annual reports in computational chemistry*, volume 4, pages 217–241. Elsevier, 2008.

Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015.

Billy Chang, Uwe Kruger, Rafal Kustra, and Junping Zhang. Canonical correlation analysis based on hilbert-schmidt independence criterion and centered kernel target alignment. In *International Conference on Machine Learning*, pages 316–324, 2013.

Anna Cichonska, Tapio Pahikkala, Sandor Szedmak, Heli Julkunen, Antti Airola, Markus Heinonen, Tero Aittokallio, and Juho Rousu. Learning with multiple pairwise kernels for drug bioactivity prediction. *Bioinformatics*, 34(13):i509–i518, 2018.

Anna Cichońska et al. Machine learning for systems pharmacology. 2018.

James C Costello, Laura M Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P Menden, Nicholas J Wang, Mukesh Bansal, Petteri Hintsanen, Suleiman A Khan, John-Patrick Mpindi, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, 32(12):1202, 2014.

Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280, 2002.

Hanna Eckert and Jürgen Bajorath. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug discovery today*, 12(5-6):225–233, 2007.

Sean Ekins, Jordi Mestres, and Bernard Testa. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *British journal of pharmacology*, 152(1):9–20, 2007.

Manuel AR Ferreira and Shaun M Purcell. A multivariate test of association. *Bioinformatics*, 25(1):132–133, 2008.

Seth Flaxman, Dino Sejdinovic, John P Cunningham, and Sarah Filippi. Bayesian learning of kernel embeddings. *arXiv preprint arXiv:1603.02160*, 2016.

Matteo Floris, Alberto Manganaro, Orazio Nicolotti, Ricardo Medda, Giuseppe Felice Mangiatordi, and Emilio Benfenati. A generalizable definition of chemical similarity for read-across. *Journal of cheminformatics*, 6(1):39, 2014.

Mathew J Garnett and Ultan McDermott. The evolving role of cancer cell line-based screens to define the impact of cancer genomes on drug response. *Current opinion in genetics & development*, 24:114–119, 2014.

Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, Chris D Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I Richard Thompson, Xi Luo, Jorge Soares, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570, 2012.

Johann Gasteiger and Thomas Engel. *Chemoinformatics: a textbook*. John Wiley & Sons, 2006.

Gene H Golub and Hongyuan Zha. The canonical correlations of matrix pairs and their numerical computation. In *Linear algebra for signal processing*, pages 27–49. Springer, 1995.

Andrew Goodspeed, Laura M Heiser, Joe W Gray, and James C Costello. Tumor-derived cell lines as molecular models of cancer pharmacogenomics. *Molecular Cancer Research*, 14(1):3–13, 2016.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005a.

Arthur Gretton, Alexander J Smola, Olivier Bousquet, Ralf Herbrich, Andrei Belitski, Mark Augath, Yusuke Murayama, Jon Pauls, Bernhard Schölkopf, and Nikos K Logothetis. Kernel constrained covariance for dependence measurement. In *AISTATS*, volume 10, pages 112–119, 2005b.

Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *Advances in neural information processing systems*, pages 585–592, 2008.

Rajarshi Guha et al. Chemical informatics functionality in r. *J Stat Softw*, 18(5):1–16, 2007.

Marc Hafner, Mario Niepel, Mirra Chung, and Peter K Sorger. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nature methods*, 13(6):521, 2016.

Lowell H Hall and Lemont B Kier. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *Journal of Chemical Information and Computer Sciences*, 35(6):1039–1045, 1995.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16 (12):2639–2664, 2004.

MJR Healy. A rotation method for computing canonical correlations. *Mathematics of Computation*, 11(58):83–86, 1957.

Swen Hoelder, Paul A Clarke, and Paul Workman. Discovery of small molecule cancer drugs: successes, challenges and opportunities. *Molecular oncology*, 6(2):155–176, 2012.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

Michael Inouye, Samuli Ripatti, Johannes Kettunen, Leo-Pekka Lyytikäinen, Niku Oksala, Pirkka-Pekka Laurila, Antti J Kangas, Pasi Soininen, Markku J Savolainen, Jorma Viikari, et al. Novel loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. *PLoS genetics*, 8(8):e1002907, 2012.

Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3): 740–754, 2016.

In Sock Jang, Elias Chaibub Neto, Justin Guinney, Stephen H Friend, and Adam A Margolin. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. In *Biocomputing 2014*, pages 63–74. World Scientific, 2014.

Gerald Karp. *Cell and molecular biology: concepts and experiments*. John Wiley & Sons, 2009.

MACCS Structural Keys. Symyx software: San ramon. *CA, US*, 2005.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109, 2018.

George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

Justin Klekota and Frederick P Roth. Chemical substructures that enrich for biological activity. *Bioinformatics*, 24(21):2518–2525, 2008.

Damjan Krstajic, Ljubomir J Buturovic, David E Leahy, and Simon Thomas. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, 6(1):10, 2014.

Simon P Langdon. *Cancer cell culture*. Springer, 2010.

Antonio Lavecchia. Machine-learning approaches in drug discovery: methods and applications. *Drug discovery today*, 20(3):318–331, 2015.

Andrew R Leach and Valerie J Gillet. *An introduction to chemoinformatics*. Springer Science & Business Media, 2007.

Jin-Ku Lee, Zhaoqi Liu, Jason K Sa, Sang Shin, Jiguang Wang, Mykola Bordyuh, Hee Jin Cho, Oliver Elliott, Timothy Chu, Seung Won Choi, et al. Pharmacogenomic landscape of patient-derived tumor cells informs precision oncology therapy. *Nature genetics*, 50(10):1399, 2018.

Ying Liu. A comparative study on feature selection methods for drug discovery. *Journal of chemical information and computer sciences*, 44(5):1823–1828, 2004.

Shuangge Ma and Jian Huang. Penalized feature selection and classification in bioinformatics. *Briefings in bioinformatics*, 9(5):392–403, 2008.

Lester W Mackey. Deflation methods for sparse pca. In *Advances in neural information processing systems*, pages 1017–1024, 2009.

Pekka Marttinen, Jussi Gillberg, Aki Havulinna, Jukka Corander, and Samuel Kaski. Genome-wide association studies with high-dimensional phenotypes. *Statistical applications in genetics and molecular biology*, 12(4):413–431, 2013.

Ultan McDermott and Jeff Settleman. Personalized cancer therapy with selective kinase inhibitors: an emerging paradigm in medical oncology. *Journal of Clinical Oncology*, 27(33):5650–5659, 2009.

Thomas Melzer, Michael Reiter, and Horst Bischof. Nonlinear feature extraction using generalized canonical correlation analysis. In *International Conference on Artificial Neural Networks*, pages 353–360. Springer, 2001.

Michael P Menden, Francesco Iorio, Mathew Garnett, Ultan McDermott, Cyril H Benes, Pedro J Ballester, and Julio Saez-Rodriguez. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS one*, 8(4): e61318, 2013.

Mario Niepel, Marc Hafner, Mirra Chung, and Peter K Sorger. Measuring cancer drug sensitivity and resistance in cultured cells. *Current protocols in chemical biology*, 9(2): 55–74, 2017.

Nifang Niu and Liewei Wang. In vitro human cell line models to predict clinical response to anticancer drugs. *Pharmacogenomics*, 16(3):273–285, 2015.

Tapio Pahikkala and Antti Airola. Rlscore: regularized least-squares learners. *The Journal of Machine Learning Research*, 17(1):7803–7807, 2016.

Tapio Pahikkala, Antti Airola, Michiel Stock, Bernard De Baets, and Willem Waegeman. Efficient regularized least-squares algorithms for conditional ranking on relational data. *Machine Learning*, 93(2-3):321–356, 2013.

Tapio Pahikkala, Antti Airola, Sami Pietilä, Sushil Shakyawar, Agnieszka Szwajda, Jing Tang, and Tero Aittokallio. Toward more realistic drug–target interaction predictions. *Briefings in bioinformatics*, 16(2):325–337, 2014.

Albert C Pan, David W Borhani, Ron O Dror, and David E Shaw. Molecular determinants of drug–receptor binding kinetics. *Drug discovery today*, 18(13-14):667–673, 2013.

Elena Parkhomenko, David Tritchler, and Joseph Beyene. Genome-wide sparse canonical correlation of gene expression with genotypes. In *BMC proceedings*, volume 1, page S119. BioMed Central, 2007.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/. ISBN 3-900051-07-0.

Nicholas Rhodes, Peter Willett, James B Dunbar, and Christine Humblet. Bit-string methods for selective compound acquisition. *Journal of chemical information and computer sciences*, 40(2):210–214, 2000.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. 1998.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

Holger Schwender. *Statistical analysis of genotype and gene expression data*. PhD thesis, 2007.

Holger Schwender, Sylvia Rabstein, and Katja Ickstadt. Do you speak genomish? *Chance*, 19(3):3–8, 2006.

Thermo Fisher Scientific. Cell culture basics handbook. *UK: Gibco*, 2015.

JL Sebaugh. Guidelines for accurate ec50/ic50 estimation. *Pharmaceutical statistics*, 10 (2):128–134, 2011.

John Shawe-Taylor, Nello Cristianini, et al. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

Clara S Tang and Manuel AR Ferreira. A gene-based test of association using canonical correlation analysis. *Bioinformatics*, 28(6):845–850, 2012.

Bruce Thompson. Canonical correlation analysis. *Encyclopedia of statistics in behavioral science*, 2005.

Viivi Uurtio, Sahely Bhadra, and Juho Rousu. Sparse non-linear cca through hilbert-schmidt independence criterion. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1278–1283. IEEE, 2018a.

Viivi Uurtio, João M Monteiro, Jaz Kandola, John Shawe-Taylor, Delmiro Fernandez-Reyes, and Juho Rousu. A tutorial on canonical correlation methods. *ACM Computing Surveys (CSUR)*, 50(6):95, 2018b.

Johan Van Meerloo, Gertjan JL Kaspers, and Jacqueline Cloos. Cell sensitivity assays: the mtt assay. In *Cancer cell culture*, pages 237–245. Springer, 2011.

Hrishikesh D Vinod. Canonical ridge and econometrics of joint production. *Journal of econometrics*, 4(2):147–166, 1976.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Peter Willett. Similarity-based virtual screening using 2d fingerprints. *Drug discovery today*, 11(23-24):1046–1053, 2006.

Egon L Willighagen, John W Mayfield, Jonathan Alvarsson, Arvid Berg, Lars Carlsson, Nina Jeliazkova, Stefan Kuhn, Tomáš Pluskal, Miquel Rojas-Chertó, Ola Spjuth, et al. The chemistry development kit (cdk) v2. 0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of cheminformatics*, 9(1):33, 2017.

Jun Xu and Arnold Hagler. Chemoinformatics and drug discovery. *Molecules*, 7(8): 566–600, 2002.

Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson, et al. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41(D1):D955–D961, 2012.